A Methodology for Pre-Post Intervention Studies: An application for a Nutritional Case Study

Beatriz Sevilla-Villanueva

Dept. of Computer Science

Universitat Politècnica de Catalunya - BarcelonaTECH

Supervisors: Dr. Miquel Sànchez-Marrè
Dr. Karina Gibert

Ph.D. programme in Artificial Intelligence
A Ph.D. thesis submitted for the degree of Doctor

2015

To my Father and to the memory of my Mother

Acknowledgments

This Ph.D. thesis work has been partially funded by TexWIN (CP-FP 246193-2) "Textile Work Intelligence by closed-loop control of product and process quality in the Textile Industry" and Diet4You (TIN2014-60557-R) "Un recomendador inteligente de dietas personalizadas desde la observación integrada de las características, hábitos y condiciones de las personas".

First, I would like to thank my mentors both Miquel Sànchez-Marrè and Karina Gibert who have given me all of their support during these years. It has been a long way with laughter and tears and I truly want to say *Thank you, I couldn't do this without you.* Also, I already said to them that I am really grateful because they supported my decision of facing the topic of nutritional genomics that it was a stranger for all of us.

Of course, I appreciate the support of my colleagues who every day have been suffered my joys and despairs over these last years. Thank you Dario, Ignasi, Sergio, Luis, Arturo, Sofia, Anna, Cristian and Jonathan.

Also, I would like to thank Maria Isabel Covas and Montserrat Fito from the IMIM group by providing the data of this work, the time for explaining this data, for all their efforts to follow the development of this thesis and for their feedback.

In addition, I would like to thank all the people who devote its life to the investigation. Specially, to those people who make possible the advances in the prevention and cure of those diseases that cause so much suffering.

Henceforth, my personal acknowledgments are written in Spanish (my mother language) because I would like that my relatives can read it.

Ha sido un largo proceso y han formado parte en él muchas personas a las que debo darles las gracias.

En primer lugar, quiero agradecerle a mi padre y mi hermana por estar siempre ahí y ayudarme en todas las decisiones que he tomado. A mi sobrino Víctor, que pase lo que pase, siempre hace que sonría y que haya más luz en mi vida.

Por supuesto, agradezco a mis amigas Zoraida, Pury, Sol, Laura, Patri y Ale que siempre tiene un momento para mí. Ellas son, sin duda, una de las partes más importantes de mi vida y parte de la familia que se puede escoger. Les agradezco eternamente su fe en mí, todos los ratos que paso con ellas haciendo de este mundo un sitio más acogedor y su amor y apoyo incondicional.

También, quiero agradecer a mi pareja Jonathan la gran fuerza que me dió cuando no se veía luz al final del túnel. Y que, en las últimas semanas, ha usado su mejor arma, la paciencia, para aguantar la presión que supone el finalizar una tesis.

No quisiera dejar de agradecer a Vitoria quien hace que mi mente esté más equilibrada por su gran apoyo constante. Ella me ha dado ánimos y fuerza para continuar, sobretodo estos últimos meses que han sido de presión máxima por las fechas de entrega impuestas.

Agradezco a Ingrid todas las risas que compartimos al inicio de la esta tesis y los sandwich que me preparaba cuando llegaba agotada. También, me hubiese gustado poder agradecer a Javi todas las conversaciones trascendentales que tuvimos por la misma época, incluso algunas llegaban a ser buenas ideas y todo.

Por último, quisiera darle las gracias a Alex por todo el cariño que me ha dado, la tranquilidad en ciertos momentos y por hacerme sentir una persona única.

A pesar de ser una tesis en nutrición, tengo que desagradecer que el desarrollo de ésta ha provocado el empeoramiento de mis hábitos alimenticios. Sin embargo, ahora tengo mÃ¡s conocimientos para reconducirlos.

Me gustaría acabar añadiendo que, espero que esta tesis y su continuación sirva para aportar mi granito de arena a la lucha contra todas esas enfermedades que podrían ser evitadas o, al menos, retrasadas. Y que, sin embargo, a día de hoy, se llevan a tantas personas dejando un hueco irremplazable en sus seres queridos.

Contents

List of Figures

List of Tables

Abstract

Nowadays, it is widely accepted that many of modern lifestyle habits such as diet and exercise, together with genetics, play an important role in the development of many diseases such as cardiovascular disease, diabetes, cancer, obesity and so on.

Thanks to the great advances in genetics in recent years, nutritional genomic science has emerged to personalize an individual's diet based on the particular needs of each person, aiming to maintain health and prevent disease.

The concept of personalized diet has been very fashionable in the last few years. Nevertheless, there is still no complete knowledge base that includes all the effects of nutrients on all existing organisms.

In order to study the relationship between genes, nutrients and health, classical clinical trials are often designed and pre-post dietary intervention studies are made. By means of these studies, and after following a certain dietary pattern for a period of time, one can compare the individuals' initial health status against their final health status, and observe the different reactions.

These studies have a high complexity due to several factors: i) they mix a large number of attributes for describing the individuals; ii) these attributes have different origins: anthropometric, sociodemographic, lifestyle habits, clinical tests, blood tests, urine tests and genetic tests. All of these also follow different patterns of influence on each individual; iii) the relationship between food and health is complex and bidirectional: it depends on the genes, the interaction between nutrients found in different types of food, and on multiple external factors, like individuals lifestyle.

This equation turns more difficult when the health condition and disease of every individual are included in the studies.

Currently, there is no methodology for analyzing problems with this structure and most of the results are reduced to classical statistical tests assessing two factors. In practice, even when in most cases an important number of characteristics (sometimes several hundreds) are measured, the multivariate interactions among all factors are rarely analyzed. The classical approach is mainly based on considering numerical attributes, compute differences and the search for a model that will relate the dietary intervention with those differences. However, this approach is not powerful enough to clearly find the factors associated to the dietary intervention in order to produce sufficiently reliable predictive models, or to assess

the expected effects of the dietary intervention on a particular individual. In some of these studies it has been found that not all attributes play the same role.

The main idea of this thesis is to build a hybrid methodology suitable to assess the effects of a dietary intervention in these pre-post studies by means of integrating statistics with Artificial Intelligence techniques in a new Data Mining methodology. Studies of this kind, have a different structure to those problems typically dealt with by AI.

Our proposal is, therefore, to provide a new methodology to extract knowledge from intervention studies when multiple interacting factors are available and influence patterns are different.

The building of a model that will incorporate the complexity of such interactions, including multiple external factors, and above all, being able to detect the subtle effects that food have on our body will be highly useful, functional and valuable for nutritionists. With this idea in mind, we propose to analyze the effect of the intervention locally in each of the different types of individuals that are involved in the study.

Integrative Multiview Clustering methodology is proposed in order to first model the types of persons that are involved in the study, finding natural clusters of persons according to their characteristics before and after the intervention. Dietary patterns of each group are also identified. The *Trajectory analysis* between the intervention, the initial and final state of individuals allows for the building of a finer partition. Then, both the *adherence* to the intervention and the *effect* of the intervention are locally analyzed in each of these new groups. In this analysis, the relationships between changes in biometric parameters, health condition as well as gene expressions, with respect to changes in prescribed diet by the intervention, are analyzed and the effect of dietary pattern is studied.

In addition to the complexity of analyzing this kind of data, a second problem arises due to the interdisciplinarity of this study: the transmission of the results to the experts. This thesis is also focused on contributing to the *automatic interpretation of results* that provides new knowledge useful to support further decision making. This part of interpretation is intended to contribute in the field of automatic results interpretation that has not been exploited and which, however, is crucial for the effective knowledge transfer between professionals from different disciplines such as, in this case, artificial intelligence and nutrition.

In the context of clustering, which is among the most popular data mining methods, interpretation provides a proper understanding of the essence of the classes obtained. In this thesis, a mixed methodology for cluster interpretation is proposed based on a mixture of interpretation-oriented visualization tools and significance tests. In addition, the limitations of some proposals are analyzed and a final *interpretation methodology that deals with nested partitions* is proposed to guarantee consistent interpretation with possible future refinements of a current partition.

The performance and applicability of this proposal is evaluated on a real data set coming from a particular pre-post study about the Mediterranean diet and olive oil effect, considering the characteristics of the persons involved, their health condition, genetic information and habits like physical exercise.

Experts assessed the reliability and utility of the proposal. They outlined that the proposed methodology is useful and finds out difficult patterns and behaviors of the individuals, regarding the intervention. The results provided by the methodology (profiles, trajectory map, adherence and effect analysis, visual tools, etc.) have shown to be useful to the experts for analyzing the pre-post intervention studies, and for making decisions based on that analysis.

Chapter 1

Introduction

Since the human genome was first sequenced in 2003 one may say that we are in the "genomic revolution" era. One of the next goals is to know what the genes do, particularly those involved in human diseases. Despite a substantial progress has been made in identifying genes function, this problem remains still unsolved.

In order to study how the genes work, interactions between environmental factors and genes must be discovered, as the behavior of the genes depends on these environmental factors. The concept "environmental factors" has a wide scope, ranging from characteristics of the place where an individual lives to the diet, exercise, emotional state or living habits that this individual may have such as smoking.

The study of the relationships between human genome, nutrition and health is called *Nutritional Genomics*. Within this global concept of nutritional genomics, two terms are used: *nutrigenomics* and *nutrigenetics* [Ordovas and Mooser, 2007]. *Nutrigenomics* studies the effect of nutrients on health through altering genome, proteome, metabolome and the resulting changes in physiology [Panagiotou and Nielsen, 2009] whereas *Nutrigenetics* studies how individual differences in genes influence the body's response to diet and nutrition.

These disciplines aim to answer *why similar diets have different effect on diverse apparently similar persons?*. For answering this question the relationship between nutrients and their function in the different metabolisms should be understood, and it seem that genes have responsibilities in this matters.

1.1 Pre-Post Intervention Studies

In order to study the relationship between genes, nutrients and health, classical clinical trials are often designed and pre-post dietary intervention studies are made. By means of these studies one can compare the individuals' initial state against their final state after following a certain dietary pattern for some period, thus observing the different reactions in them. These studies have a high complexity due to the fact that they mix a large number of attributes for describing the individuals. These attributes have different origins such as anthropometric, sociodemographic, habits, clinical tests, blood tests, urine tests and genetic tests. In addition, the interactions between the diet or other environmental factors with genes are also complex. This complexity is due to the fact that one nutrient can affect the function of several genes and, on the other hand, one gene can be affected by several nutrients. This equation turns more difficult when pathological states are included in the studies.

Currently, the most used techniques in this kind of studies are traditional, and quite basic statistics, as it will be described in the *State of the Art*, which find associations relating a small number of nutrients and genes.

Therefore, there are two challenges in this area: to find interactions involving several genes and several factors, and finding the genes involved in the hidden effect of the nutrition.

We are proposing a methodology that can help this type of studies. This methodology will include AI techniques in order to obtain a better information extraction. Studies of this kind have a different structure than typical problems which AI deals with.

Most of dietary intervention pre-post studies have a similar structure: examples of these studies can be found in projects such the *Framingham Heart Study* [Framingham Heart Study, 2015] or PREDIMED [PREDIMED, 2015].

The *Framingham Heart Study* is one of the most famous epidemiological study to identify the common factors or characteristics that contribute to cardiovascular diseases.

The PREDIMED project is a long-term nutritional intervention study aimed to assess the efficacy of the Mediterranean diet in the primary prevention of cardiovascular diseases.

From these projects multiple dietary intervention pre-post studies have been published (see Chapter 2).

Generally, a target population is defined and conveniently sampled. The selected involved individuals are assigned to a dietary intervention. Then, some target parameters of participants are measured *before* the dietary intervention and compared with their values *after* the dietary intervention.

Figure 1.1 shows the basic structure of this kind of studies. So, the following 5 steps can be identified:

1. Selection of sample (inclusion criteria). These criteria depend on the study goals.

2. The individuals are split in different groups depending on the number of different dietary interventions (usually is only one) and the different interventions are assigned to these groups. Normally there is an extra group with no dietary intervention assigned which constitutes the *control* group.

3. Target data from the individuals are gathered before the intervention.

4. Target data from the individuals are gathered after the intervention.

5. Pre-post analysis.

The common target of this kind of studies is to find whether the dietary intervention has an effect over the individuals. This target is derived from nutritional genomics' goal, which seeks the understanding of how nutrition influences metabolism. Within this goal different purposes may coexist. For instance, depending on whether the search is within a healthy population or in a population with some disease. When dealing with healthy individuals, the way how diet affects them is often regarded both at a metabolic level and at a genetic level. This is done so to identify which genes are sensitive to these nutrients. Also, to observe which genes are more active (more expressed) in the presence of these diet guidelines.

Nevertheless, regarding individuals with some disease -where differences between healthy and not healthy are compared, or just unhealthy- different purpose can arise. Finding the genes that are related with the studied diseases or analyzing how different the interaction between nutrient and gene is in pathological states than in healthy state. Specially interesting is to compare individuals against some diseases with individuals that have high risk to suffer that disease but it is not currently present. In that way, it is possible to identify nutrients which can help to prevent the diseases. Besides, to identify certain nutrients that help to reduce the effect of such disease.

Data included in such studies come from surveys and medical tests which have been made to the involved subjects. In most of cases an important number of characteristics (attributes sometimes several hundreds) are measured. *The classical approach is to consider only numerical attributes, compute differences and search for a model relating the diet with those differences.* However, this approach is not powerful enough to clearly find the factors associated to diet. In some of these studies it has been found out that not all attributes play the same role, as it will be explained later.

This is a complex problem due to many reasons. Among them, the effect need not be reflected in just one gene. So, multiple factors associations are searched. Moreover, this complexity comes from the relatively small effect of dietary interventions on physiological parameters. Similarly, nutrition effects on gene-expression patterns are also hard to detect [Afman and Müller, 2006].

The main idea of this thesis is to discover the effect of the intervention introducing techniques from Artificial Intelligence field. Concretely, we propose to use an Integrative Multiview Clustering methodology to find specific types of individuals according to their characteristics, and also according to the dietary intervention, at the beginning and at the end of the intervention. Afterwards, the effect of the intervention is locally analyzed in each crossed cluster found and comparing both initial and final states.

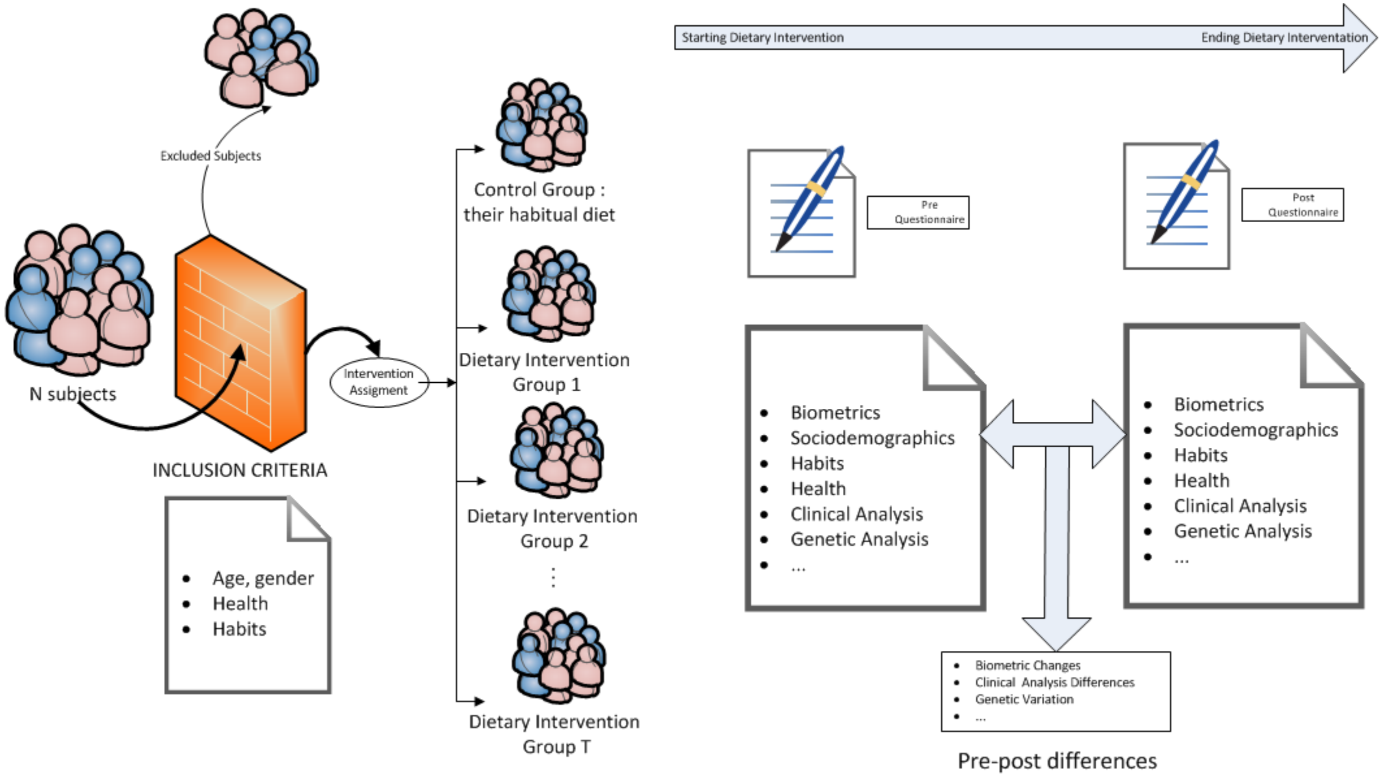


Figure 1.1: Structure of an Intervention Pre-Post Study

1.2 Clustering

In this thesis, clustering methods are proposed in order to model the types of persons that are involved in the study, finding natural clusters of persons according to their characteristics. This characterization is used for the local study of the diet effect over the found profiles.

Besides of the complexity of analyzing this kind of data, a second problem arise due to the interdisciplinarity of this study: the transmission of the results to the experts.

When transmitting the results of data mining analysis to the end user, it is important that the results are completely understood [Gibert and Tormos, 2014]. A process of results interpretation provides new knowledge that can be used to support further decision making. In fact, interpretation is crucial for proper knowledge transfer, especially to experts from other disciplines.

In the context of clustering, which is among the most popular data mining methods, interpretation provides a proper understanding of the essence of the classes obtained, either if they have been automatically discovered or expert-based. Thus, a good characterization of the classes requires proper interpretation tools for post-processing the clusters themselves [Gibert et al., 2013, Gibert and Conti, 2014].

In general, the objective of clustering is to generate a set of different classes[1] that group similar individuals in the same class. Therefore, individuals of the same class can be described by common characteristics, and some of these characteristics are expected to be different from other classes. *Cluster Interpretation* is a post-process of finding the common and distinctive characteristics of every class, and creating the corresponding profiles. Most of the work in the literature about analyzing clustering results is focused on cluster validity, and validity indexes are used to provide the structural validity of classes [Arbelaitz et al., 2013]. Structural validity does not necessarily ensure the usefulness of clustering, as meaningfulness is also key to guaranteeing decision-making support. However, cluster interpretation is still an open issue from the methodological point of view, and with respect to the automation of this process, although there are few works on these topics. Some proposals [Gibert et al., 2008a] based on the analysis of conditional distributions among classes are already available in this line. In most of the works, concept induction [Gibert, 2014], or statistical tests [Rudolph and Gibert, 2014, Sevilla-Villanueva et al., 2013], are used to identify which variables behave differently in some of the classes.

[1] Class and Cluster: In the literature, there is not a common agreed definition and use of both terms. For many authors, both terms are synonyms. For other, they have some light differences. Usually, the term "class" is used more frequently in mathematical texts, and the term "cluster" is used more frequently in computer science texts.

Here in this document, the term cluster is used whenever the term is more related to a clustering process. The usual term used is class, with an implicit sense that a class is a cluster with some interpretation of its elements (semantic addition).

An existing partition of a set of objects sometimes becomes enriched in a second moment when a refinement of the original partition is made, and a new nested partition is provided. This might correspond to a partition of higher granularity than the first, or to the incorporation of new information in the clustering process, or to a combination of two or more previous partitions of higher granularity in a Cartesian product [Sevilla-Villanueva et al., 2014] that produces nested classes considering criteria used in all the original partitions. The referenced interpretation proposals do not guarantee robust behavior in this case, and might produce contradictory interpretations, as variables that were relevant for characterizing parent classes may disappear from the description of the refined partition.

In this thesis, the limitations of the some proposals are analyzed and a final interpretation methodology is proposed to guarantee consistent interpretation with future refinements of a current partition. A mixed methodology for cluster interpretation is proposed that is based on a mixture of interpretation-oriented visualization tools and significance tests. *Class Panel Graphs* [Gibert and Sonicki, 1999a] are used for the visualization, whereas, *Test-Value* [Lebart et al., 2000] have been imported from the factorial analysis field of multivariate statistics. The use of Test-Value for cluster interpretation was introduced in [Sevilla-Villanueva et al., 2013]. In this thesis, a contribution is introduced on how Test-Value needs to be generalized to better identify the class characteristics. A modification of the interpretation methodology itself is also introduced to guarantee the consistency of an interpretation with future refinements in a new nested partition. Sensitive Analysismethods have been used for this purpose.

In this thesis, contributions in Cluster Interpretation are introduced: the use of Test-Value for cluster interpretation and on how Test-Value needs to be generalized to better identify the class characteristics. The analysis of consistent interpretation when nested partitions coexist.

1.3 Goals

1. Formalization of a methodology to allow analysis of complex pre-post dietary intervention studies.

2. Proposal of an integrative multiview clustering method for a reliable analysis if individuals behavior in pre-post intervention studies.

3. Proposal of a mixed methodology for cluster interpretation

   (a) The use of *Test-Value* for identifying the characteristics of a cluster

   (b) The generalization of the *Test-Value*

   (c) Solving the consistency problems of nested partitions

4. Proposal of the Trajectory Map and the characterization of the resulting trajectories to analyze the adherence to the intervention and the effects of this intervention.

1.4 Application

The performance and applicability of this proposal is evaluated on a real data set of baseline characteristics, diet habits, and levels of physical activity of a sample of the general population. The proposed class interpretation process can contribute to identifying standard nutritional patterns in the general population and their association with health conditions and physical activity habits, which is aligned with new preventative 'healthy life-style' policies for a better health condition, especially in the long term and aging. Obtaining a clear interpretation of the nutritional patterns in the population will permit the establishment of dietary guidelines to increase the health of persons and reduce public health costs in the long term.

In this work data from a particular study of olive oil effect will be used as a case study to show the benefits of the presented proposal.

Data comes from an starting and initially not funded collaboration between the research team leaded by María Isabel Covas and Montserrat Fitó, Cardiovascular Risk and Nutrition Research Group at Institut Municipal d'Investigació Mèdica in Barcelona and Doctors Miquel Sànchez Marrè and Karina Gibert from Knowledge Engineering and Machine Learning Group at UPC. Currently, a national project *Diet4You* (TIN 2014-60557-R) has been approved.

1.5 Schema of this document

The following chapters are organized as follows. In Chapter 2, first of all, there is a brief description about nutritional genomics and what is its current state. Second, the pre-post studies are explained and a review of the these studies is provided from a methodological point of view. A clustering section is included because we propose to use clustering for the analysis of the pre-post studies. Finally, since this proposal has a strong component in cluster interpretation, we include a section about this topic.

Chapter 3 includes the formalization of the PhD thesis problem and presents the proposed methodology to solve it.

In chapter 4 the tools used to develop this thesis are described.

The application of the proposed methodology on the case study is detailed in Chapter 5. The case study is described in Section 5.1. Section 5.2 gives details of the complete preprocessing process. Section 5.4 shows the results of the clustering of the initial states of the persons. Section 5.5 shows the resulting partition of the final state of the individuals. In Section 5.7, the evaluation of the adherence to the intervention is performed. Then, the effects of the intervention are analyzed by comparing both states of individuals in Section 5.8.

Chapter 6 presents the justification of the different parts of the proposed methodology. The evaluation of the Integrative Multiview Clustering and the selection of Ward's method. Also, the Cluster Interpretation Methodology is evaluated comparing it with classical approach and the benefits of the methodology are presented.

Chapter 7 contains some related research including an alternative local version of *Test-Value* , the use of the frequent itemsets for cluster interpretation and the analysis of the most common cluster validity indexes.

Finally, Chapter 8 concludes the research of this PhD Thesis including the conclusions, contributions and future work.

In addition, the related publications to this work are in Chapter 9.

At the end of the document, the glossary and two appendices are included. The first Appendix A is a description of all the attributes included in our case study and the second Appendix B contains additional information about the interpretation of the resulting partitions.

Chapter 2

State of Art

The main goal of this thesis is to analyze a nutritional genomic pre-post intervention study using Artificial Intelligence. Nevertheless, nowadays, there is not enough literature found about this approach. For this reason, in this chapter, the different topics are separately explained.

First, the concept of Nutritional Genomics is explained in Section 2.1. Then, the pre-post studies in general are explained including some references of how tackle this type of studies and the works that we have found in nutritional intervention studies.

Since our approach is based on Clustering, Section 2.3 copes with this topic and some references to works using clustering with nutritional data or genetic data are contemplated.

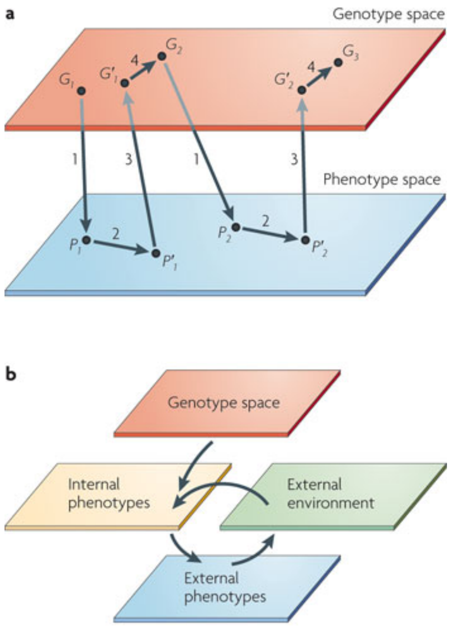
Also, two sections are devoted to cluster validation (Section 2.5) and cluster interpretation (Section 2.6). Since, this thesis contains an important component in cluster interpretation, the subsequent sections contain the concepts that are directly related with this topic such as the statistical tests or the Class Panel Graphs.

2.1 Nutritional Genomics

Nutritional genomics has a tremendous potential to change the future of dietary guidelines and personal recommendations. Nutritional genomics represents the application of systems biology to nutritional research.

Prior to start the description, the idea of genotype-phenotype relationship has to be understood. Figure 2.1 from [Houle et al., 2010] illustrates the relationship between both concepts. Phenotype is any trait that can be observable. A more general description is whatever is changed in the organism whenever a gene's function is altered [Benfey and Mitchell-Olds, 2008]. The phenotype is also sometimes defined as an equation depending on the genotype and the environmental factors.

As it was mentioned in Chapter 1, Nutritional Genomics studies the relationship between human genome, nutrition and health. The two main disciplines that are included in nutritional genomics are depicted in Figure 2.2. Briefly, these disciplines can be described as following:



Nature Reviews | Genetics

Figure 2.1: Phenotype - Genotype - Environmental Factors

Nutrigenomics: studies the effect of nutrients on health through altering genome, proteome, metabolome and the resulting changes in physiology [Panagiotou and Nielsen, 2009]. In Figure 2.2 the different variants depending on which level is analyzed of the gene expression are shown.

- Nutritranscriptomics: valuation of the transcription of DNA and analysis of changes in the gene expression measured to mRNA level through different nutritional conditions. In other words, it is the study of mRNA expression levels in a single cell or in a population of biological cells for a given set of nutritional conditions.

- Nutriproteomics: large-scale analysis of the structure and function of proteins as well as of protein-protein interactions in a cell.
- Nutrimetabolomics: the measurement of all metabolites to access the complete metabolic response of an organism to a nutritional stimulus.

Nutrigenetics: studies the effect of genetic variations on the interaction between diet and health with implications to susceptible subgroups. More specifically, nutrigenetics studies how individual differences in genes influence the body's response to a specific dietary pattern, a functional food, or a supplement for a specific health outcome.

These specific fields of nutritional genomics are emerging as important new research areas. The reason is that it is becoming increasingly evident that damage to genome is the most fundamental disease. A risk for developing a disease increases with DNA damage which - in turn - is dependent on nutritional status and an optimal dietary intake. Also, tissue concentration of nutrients for prevention of genome damage is dependent on genetic polymorphisms which alter function of genes. These genes are involved directly or indirectly in the uptake and metabolism of nutrients [Fenech, 2008].

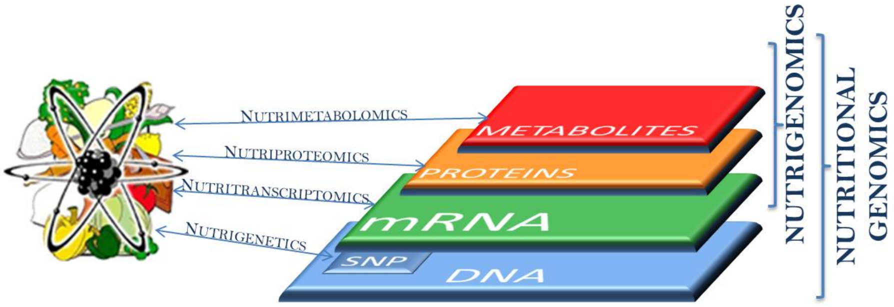


Figure 2.2: Nutritional Genomics

These disciplines study diet-gene-disease interactions and aim to promote health and disease prevention. They are based on the idea that everything ingested into a person's body affects the genome of the individual and therefore, both genes and nutrients modify the same metabolic processes.

In the following paragraphs, the complexity of such studies is justified.

It is a complex relationship due to several reasons. First, because depending on how the genes are expressed nutrients are metabolized in a particular way. Also, depending on the nutritional status of the individual genes may be differently expressed. In addition, if the individual is suffering a disease the behavior of these interactions vary. Moreover, depending on how these interactions are, the development of the disease can be boosted.

Besides, most chronic diseases, such as cardiovascular diseases, metabolic syndrome and cancer are multifactorial disorders caused by multiple genetic and environmental factors. The diseases which have a genetic basis can be classified depending on whether there is one gene involved (monogenic) or more than one gene (polygenic). The genes involved in many of these polygenic diseases are not known yet. Currently, most results have been dealing with monogenic diseases as it is an even simpler problem remains complex.

Also, the effect of dietary interventions on physiological parameters are relatively small. Similarly, effects of nutrition on gene-expression patterns are also difficult to be detected [Afman and Müller, 2006].

Note that the different disciplines can be complementary in two senses: first, conclusions from different disciplines can be contrasted giving more robustness to the common results. Second, results of one study can be used to select genes biomarkers for other studies, thus reducing their search space. For instance, a study that simultaneously identified a mechanism for the regulation of sterol uptake in the intestine that is the basis for sitosterolemia (a genetic disorder characterized by hyperabsorption of dietary sterols leading to hypercholesterolemia with a high risk of developing atherosclerosis) [Berge et al., 2000]. This study found the mutation responsible of the uncontrolled hyperabsorption of dietary sterols from a previous nutrigenomic study where the gene responsible was identified. This previous study used mice.

2.1.1 Genomic Data

The structure that today is known as DNA (Deoxyribonucleic acid) was described by Watson and Crick in 1953 [J.D. and F.H.C, 1953] thanks to the work of Rosalind Franklin [Franklin and Gosling, 1953, Maddox, 2003]. After this fact, ideas and new experimental approaches came together. These led to what has been called the central dogma of molecular genetics, which states that genetic information flows from DNA to RNA (Ribonucleic acid) and from there to the protein, and therefore DNA is the material that holds genetic information on hereditary information units called genes.

A gene is a sequence of DNA that contains genetic information and can influence the phenotype (observable trait) of an organism. Within a gene, the sequence of bases along a DNA strand defines a messenger RNA sequence (mRNA), which then defines one or more protein sequences. The relationship between the nucleotide sequences of genes and the amino-acid sequences of proteins is determined by the rules of translation, known collectively as the genetic code.

The gene expression is the process by means of all organism transform the coded information of DNA into RNA (transcription). In all organisms, the DNA is identical in all their cells but not all genes are expressed at the same time nor in all cells. Instead, they are expressed according to certain factors which regulates the gene expression [Balanza, 2007]. These factors are the following:

- Cell function in a specific tissue.
- Response to external stimuli (environmental factors).
- Pathological states.
- Temporal specificity depending on la stage of the life of the organism.

The Microarray technology [DeRisi et al., 1996] is the most used technique to obtain the data so as to analyses the gene expression [Lausted et al., 2004]. This technology was possible thanks to the invention of PCR (Polymerase Chain Reaction) which revolutionized the study of DNA [Robinson, 2010]. PCR was conceived by Kary Mullis and his colleagues in 1983 [Bartlett and Stirling, 2003]. PCR is an enzymatic method for amplifying specific DNA sequences. This method let the synthesis *in vitro* of a DNA sequence. There are several variants according to the different purpose such as:

- Reverse Transcription PCR (RT-PCR) that amplify DNA from RNA. This is used for gene expression profiling.
- Allele-specific PCR: diagnostic or cloning technique based on single-nucleotide variations. This is used for Single-Nucleotide Polymorphism (SNPs) genotyping.

The process of creation of a microarray (commonly known as *DNA chip* or *biochips*) consist in a collection of microscopic DNA spots attached to a solid surface. These DNA spots represent the expression levels of large numbers of genes simultaneously or the genotype of multiple regions of a genome. Each DNA spot contains a very small quantity of a specific DNA sequence (known as probes) that have been exposed to certain conditions which are the target of the study.

Subsequently, after several biological processes (hybridization, washing and developing) an image is obtained in which the luminous intensity of each one of the cells (spots) of the microarray can quantify the level of expression of each gene, under each experimental condition [Segal et al., 2003].

Finally, an array of numerical data from the pattern of intensities of each cell is generated using image analysis techniques. The result is a two-dimensional matrix of numeric data, normally encoded by real values, where one of the dimensions represents the different genes, and the other represent the conditions under which genes have been analyzed. Each values in the matrix represent, therefore, the level of expression of a given gene under some experimental condition.

An example of use of this technology is [Golub et al., 1999]. In this article they highlight the importance of standards for obtaining Microarray because their results did not work when they used microarrays from other laboratories.

2.1.2 Current Research

Advances in Nutritional Genomics are enormous in last years. Specially since the gene expression can be translated into microarrays formats which can after be analyzed (see Section 2.1.1). Thus, this genomic data can be introduced in studies such as the dietary intervention pre-post studies. Anyway, there is still a long way to reach a highly personalized diet [de Lorenzo, 2012].

DNA is object of study since several decades but it was not until 2001 when the first draft of the human genome was reported [Lander et al., 2001,Venter et al., 2001] and in 2003 when the complete human genome sequence was finished [International Human Genome Sequencing Consortium et al., 2004] in the Human Genome Project [Human Genome Project, 2015]. Nevertheless, at this time other organism's genomes such as bacterias, yeast, insects or mice were already sequenced.

Although many results have been published about human, most of the results that are recognized by official organizations such as the World Health Organization (WHO) [WHO, 2015] or the (U.S. Food and Drug Administration (FDA) [FDA, 2015] are based on monogenic diseases. For instance, phenylketonuria is a monogenic disease that the effect can be avoided with the correct diet guidelines [Mitchell et al., 2011]. Or based on interactions which includes few factors which are widely studied as smoking or trans-unsaturated fat intaking in cardiovascular diseases.

Nutritional genomics is quickly evolving. This growth has generated new organizations so as to share knowledge. Perhaps the most famous is The European Nutrigenomics Organization (NuGO) [NuGO, 2015]. NuGO evolved from a European-funded Network of Excellence and it is an Association of Universities and Research Institutes focusing on jointly developing the existing research area of nutrigenomics and nutritional systems biology. NuGO is currently expanding globally around the world. In New Zealand, *Nutrigenomics New Zealand [Nutrigenomics New Zealand, 2015]* has been consolidated. *Nutrigenomics New Zealand* is a large multi-disciplinary team from Research Institutes and a major university that has been established for the benefit of New Zealand health and the New Zealand food industry. Their research focused on building an understanding of food, food components, autoimmune diseases and gut health. Other organizations are JINGO [JINGO, 2015] and ISNN [ISNN, 2015] and CENG [CENG, 2015]

As an example of the relevance of these new disciplines is found in *The Framingham Heart Study [Framingham Heart Study, 2015].* Being one of the most famous epidemiological study which started in 1948 aims to identify the common factors or characteristics that contribute to cardiovascular disease. In last years, in order to achieve their objectives, they have included the study of the genes related to cardiovascular diseases and how the environmental factors affect to them, including diet [Millen et al., 1996, Corella et al., 2009, Jaimungal et al., 2011, Kimokoti et al., 2012, Kofler et al., 2012].

The publications both in nutrigenomics and nutrigenetics are multiple but not all reliable as it is explained in [Lee et al., 2011]. This article is a review of the gene-environmental factor interactions found in literature. They report that in many studies there is a lack of statistical significance, that many published results are contradictory and in addition, many studies are not reproducible.

In [Gibert and García, 2008] is pointed out a *need for new methodologies to analyze the efficacy of interventions when contextual factors that can interact with the intervention* itself are not all well-known. These conditions make difficult to design a correct pre-post studies as they were initially conceived and decrease the reliability of the provided results. The traditional techniques used in these studies not always extract all relevant information from data.

A big limitation is obtaining data. In general, the epidemiological studies require a high cost both in time and money. In addition, nutritional genomics requires genetic tests for every individual and they are very expensive. For these reasons, many studies include a small sample of persons. Besides, nutritional genomics is an innovative area which implies that the number of these studies is scarce and there is not open available data as in other fields. For these reasons is difficult to find data about nutritional genomics pre-post studies.

Table 2.1: Classification of the references based on the type of analyzed data and the methodology

Data	**Method**	**Target**	**References**
Nutritional Intervention	Statistical Pairwise test	Find whether the intervention prevents esophageal cancer.	[Li et al., 1993b, Li et al., 1993a, Mark et al., 1994, Taylor et al., 1994, Li et al., 1995, Zou et al., 2002]
Medical Records	Decision Trees, logistic regression, Association rules	Explain risk factors	[Chae et al., 2001, Tai and Chiu, 2009]
Microarrays	Hypothesis test	Identify relevant genes which change along the treatment.	[Mutch et al., 2005, Ordovas and Mooser, 2007]
Gene Expression & environmental factors	Statistics (t-Student test, Wilcoxon test, Leveneo's Test, linear regression , survival models)	Correlate genes with environmental factors.	[Coltell i Simon, 2004, Tai et al., 2005, Tucker et al., 2005, Farina et al., 2011]
Gene Expression & Olive oil and/or Mediterranean diet	Statistics (ANOVA, Kruskal-Wallis test, Wilcoxon,linear models, χ^2-Independence test, Cox regression)	EUROLIVE studies the olive oil effect and PREDIMED the benefits of the Mediterranean diet using atherosclerosis-related genes	[Khymenets et al., 2009, Konstantinidou et al., 2010b, Konstantinidou et al., 2010a, Camargo et al., 2011, Castañer et al., 2012, Perona et al., 2011, Salas-Salvadó et al., 2011]
Nutritional Intervention & Gene Expression in Animals	Clustering	Typify the different diets.	[Loor et al., 2011, Loor and Bionaz, 2012]
Nutritional Intervention & Gene Expression in Humans	Hierarchical Clustering	Group genes with similar changes in their expression	[Dahlman et al., 2005, Clement et al., 2004, Viguerie et al., 2005]

2.2 Pre-Post Studies

In this section, first the concept of Pre-Post study is introduced and the types of studies that exist. Then, some references of some works are explained either because they have been analyzed using AI or are dietary intervention. In the following Table 2.1, there is a resume of these publications.

A *Pre-Post* study examines whether participants in certain intervention improve or regress during the course of the intervention, and then, it attributes any such improvement or regression to the intervention. This definition is from The National Center for Technology Innovation [National Tech Center, 2013]. A pre-post study is a type of evaluation that seeks to determine whether a program or intervention had the intended causal effect on program participants. Intervention is understood in a wide sense and can be, in the health context, either a pharmacological treatment, a neurorehabilitation program or a diet, among others. There are two key components of a pre-post study design: pre-post test design and a treatment group and a control group.

A *pre-post test design* requires that data is collected on studied participants before the intervention takes place (pre), and that the same data is measured again after the intervention takes place (post). This design is considered the best way to be sure that your intervention had a causal effect.

To get the true effects of the program or intervention, it is necessary to have both a treatment group and a control group. As the name suggests, the *treatment group* receives the intervention (or is subdivided to receive different kind of interventions to be compared). By having both a group/s that received the intervention and another group that did not (*control group*), researchers control for the possibility that other factors not related to the intervention itself are responsible for the difference between the pre and post results. It is also important that both the treatment group and the control group are of adequate size to be able to determine whether an effect took place or not.

Finally, it is important to make sure that both the treatment group and the control group are statistically similar. While no two groups will ever be exactly alike, the best way to be sure that they are as close as possible is having a random assignment of the studied participants into the treatment group and control group. By randomly assigning participants, it can be ensured that any difference between the treatment group and control group is due to chance alone, and not by a selection bias.

As mentioned above, these studies are best used to address whether a program or intervention had the intended causal effect on program participants. Further, it is necessary that the program or intervention can be measured quantitatively in some fashion (through a knowledge test, observations, survey questions, etc.). The general form of a research question that such an experimental study can answer is *"What is the effect of intervention on a specific population?"*

A methodology for designing experiments was proposed by Ronald A. Fisher, in his innovative books: "The Arrangement of Field Experiments" (1926) and "The Design of Experiments" (1935).

Laws and ethical considerations preclude some carefully designed experiments with human subjects. Legal constraints are dependent on jurisdiction. Constraints may involve institutional review boards, informed consent and confidentiality [Moore and Notz, 2006].

Typically, these studies are commonly used for drug development. The data of these studies is commonly analyzed using traditional statistics (pairwise hypothesis test) or statistic regressions.

Health research is currently interested in getting scientific evidence from a wider spectrum of health related issues, other than drug effect like, neurorehabilitation, nutrition, mental health, behavior or social science. In the particular field of nutrition there are several studies that aim to study the diet effect at domestic doses. For these situations, the traditional techniques used in pre-post studies not always extract all relevant information from data [Gibert and García, 2008]. For these contexts it is difficult to ensure the real homogeneity of the groups, in the sense that many non-measured contextual factors makes almost impossible to find persons with equal potentials to respond to intervention. This is the key of case-control studies which assume that differences in the final situation of the patient receiving the intervention with respect to a similar one in the control group are attributable to the intervention itself. In [Gibert et al., 2008b] is pointed out a need to new methodologies to analyze efficacy of interventions when contextual factors that can interact with the intervention itself are not all well-known and not easy to measure or too much to be explicitly considered. These conditions make difficult a correct design of pre-post studies as they were basically conceived and decrease the reliability of the provided results.

Nowadays, it is well known that Knowledge discovery (KDD) provides a good framework to complex phenomena [Gibert et al., 2008b], as the one referred above, for getting novel and valid knowledge that can improve the background *doctrine corpus*.

Nutritional intervention using statistics

In the particular field of nutrition, several nutritional intervention trials can be found. As an example, the one conducted in Linxian, China between 1985-1991: Dysplasia Trial and General Population Trial. These trials tested the effect of multiple vitamins and minerals in the prevention of esophageal cancer in a population. Being the China population one of with the highest rate for this disease in the world. Some publications of these trials are [Li et al., 1993b,Li et al., 1993a,Mark et al., 1994,Taylor et al., 1994,Li et al., 1995,Zou et al., 2002]. As said before, basic statistical pairwise test are mainly used to conclude in this kind of studies.

Medical records using AI

In other medical fields, some references are found about studies which introduce some AI techniques for the analysis. In these works, they state that the use of data mining tools can be useful to extract knowledge from this kind of data and understand the risk of a specific disease. In [Chae et al., 2001], a study with two datasets provided by Korea Medical Insurance Corporation (KMIC) was performed. The first dataset was composed by 50% of the total beneficiaries who were randomly selected. In the second dataset, 100% of the beneficiaries with hypertension and an equal number of beneficiaries without hypertension were randomly selected. This paper compared the results obtained by a logistic regression and two different decision trees (CHIAD and C4.5), concluding that logistic regression produce a better explanation of the risk factors, but CHIAD was a better predictor of hypertension. In addition, they used association rules for providing specific information about risk factors. In [Tai and Chiu, 2009], from clinic records of enrollments of Taiwan National Health Insurance (NHI), youngsters with diagnosis of Attention deficit-hyperactivity disorder (ADHD) in 2001 were recruited as case group in this study. And all their clinic diagnoses made from 2000 to 2002, as well as comorbidity, were categorized. For comparison, fourfold non-ADHD controls were recruited from 2001s NHI enrollments on a random base but matching gender and age of cases. In this study statistical models and association rules were used to study the comorbidity of ADHD.

Microarrays using statistics

With the beginning of the Humane Genome Project in 90's, there is an increasing interest extracting knowledge from genetic data. Most of these works analyze microarray data directly. Basic hypothesis test are often used to test significant changes in gene expression [Mutch et al., 2005, Ordovas and Mooser, 2007]. This identifies a subset of relevant genes which change expression with the treatment.

Gene Expression & environmental factors

Basically, classical statistical modeling techniques and the most frequent correlation analysis (regression) are used in this type of studies. In some other works, the set of relevant genes is then introduced in a second part of the analysis by searching relationships between the genes and other non-genetic attributes of interest. Classical statistical modeling is the most frequent in this context. Like on the case of the *Framingham Heart Study*. Among their results, in [Coltell i Simon, 2004], there is a compilation of different results of one of the studies *Framingham Off Spring Study* included in the original *Framingham Heart Study*. This PhD.Thesis refers that the methods and techniques that have been used in the area are from the traditional statistics. They compare means, variance and distributions by means of the t-Student test, Wilcoxon test, Levene's Test, ANOVA or χ^2 Independence test. In some cases, for associating two numerical attributes, they use the linear regression. Two examples of published results which use these techniques are [Tai et al., 2005, Tucker et al., 2005]. Also, from the *Framingham Heart Study* is the work presented in [Farina et al., 2011]. In that case, a proportional hazards model, which is a class of survival models in statistics, is used to associate that achidonic acid and alpha-linolenic acid with a reduced risk of hip fracture in older adults.

A wider research was done over databases as *ClinicalTrials*[1] or ICTRP[2] using any combination of the following key words: *genetic, genomic, dietary intervention, clustering or data mining, nutrient, olive oil, etc.* However, it seems that the AI techniques are not too popular yet in this application field.

Gene Expression & Olive Oil

As far as our data comes from an Olive Oil Intervention study, in the following paragraphs there is compilation of some results in important trials about the benefits of the Olive Oil.

[1] ClinicalTrials.gov is a registry and results database of publicly and privately supported clinical studies of human participants conducted around the world

[2] http://www.who.int/ictrp/ : International Clinical Trials Registry Platform

In the European project EUROLIVE, several studies in humans were carried out. The objective of these studies was to obtain scientific evidence on the impact of olive oil, and its phenolic compounds, on oxidative stress and damage in several European populations. In [Covas et al., 2006] is published the results of a randomized study in order to evaluate whether the phenolic content of olive oil further benefits plasma lipid levels and lipid oxidative damage compared with monounsaturated acid content. In this study, the data was normalized with the log-transformed if were necessary. The 1-factor analysis of variance (ANOVA) or Kruskal–Wallis test were used, as appropriate, to determine differences in baseline characteristics. The Student t-test was used to determine differences in baseline characteristics between participants who did and participants who did not complete the study. Some of the resulting publications are the following:

In [Khymenets et al., 2009], the objective was to identify the genes that responds to virgin olive oil. The changes in all biochemical blood parameters were analyzed using the statistical Wilcoxon signed rank test.

The work [Konstantinidou et al., 2010b] examines if fat load induce changes in the expression of insulin sensitivity-related genes. In this study the Pearson's correlation test is used as a correlation analysis and the t-test was performed to assess gender differences.

In [Konstantinidou et al., 2009], a study was carried out with the aim to assess the gene expression changes in genes involved cardiovascular diseases on 6 healthy volunteers. The comparisons pre-post are described with descriptive statistics but there were not control groups for comparing.

In [Konstantinidou et al., 2010a], a randomized, parallel, controlled clinical trial in healthy volunteers was studied. The aim of this study was to assess whether benefits associated with Mediterranean diet and virgin olive oil consumption could be mediated through changes in the expression of atherosclerosis-related genes. The relationships between continuous attributes were measured by Spearman's rank correlation coefficient, t-test or general modeling statistics. Eventually, the ANOVA test was used for assessing difference between the control group and the dietary intervention groups.

In [Camargo et al., 2011], the effect of phenolic compost of olive oil has been studied for patients with metabolic syndrome who are persons with a high atherosclerosis risk. The differential analysis of the gene expression was under-fitting linear models. Differences between plasma concentration after intaking both types of olive oil were analyzed using Wilcoxon-paired test.

In [Castañer et al., 2012] were published the results of the randomized, crossover, controlled study. This study concluded that polyphenols could exert health benefits. In this study, the data was transformed to a normal distribution and Pearson's correlation analysis were used to calculate relationships among attributes. A paired t-test was performed to assess the effect of each intervention compared with its baseline. They used adjusted general linear mixed models so as to assess the effect of interventions.

The project PREDIMED [PREDIMED, 2015] started in October 2003 and it is a long-term nutritional intervention study. The complete title of it is "Effects of the Mediterranean diet on the primary prevention of cardiovascular diseases". It aimed to assess the efficacy of the Mediterranean diet in the primary prevention of cardiovascular diseases. It is a large randomized clinical trial of dietary intervention in persons at high risk of cardiovascular disease. The multidisciplinary team of the PREDIMED study assembles outstanding research groups involved in nutrition and cardiovascular risk in Spain. Partners were 16 groups distributed in 7 autonomous communities in Spain, which are formed by university researchers, hospital clinicians, primary care physicians, nutritionists and epidemiologists working in various public institutions. Two important publications are the following.

The work [Perona et al., 2011] shows the results of testing whether phenolic compounds can modulate the serum and very low-density lipoprotein triacylglycerol concentrations in humans. A double-blind, randomized, crossover trial was designed. To analyze the difference among the three groups One-factor analysis of variance and Kruskal-Wallis tests were used. A general linear model for measurements was used, using the Pearson's coefficient, with multiple comparisons corrected by Tukey's method, to assess differences among groups.

The work [Salas-Salvadó et al., 2011] tests the effects of two Mediterranean-diet interventions versus a low-fat diet on incidence of diabetes. The results were presented comparing three different groups. Comparisons among these groups for qualitative attributes were done with the χ^2-Independence test. Cox regression model was used to assess the relative risk of diabetes by allocation group, estimating hazard ratios and 95% confidence intervals (CI). Kaplan-Meier survival curves were plotted to estimate the probability of remaining free of diabetes during follow-up. Analyses were based on the intention-to-treat principle.

Most of these studies use *classical statistics* to see whether there is an association between one gene and one environmental factor and therefore, the studied interactions are one to one or included few selected genes or factors.

Authors are not aware of works in which clustering has been used to find more general conceptual entities to be used as inputs of global models establishing the relationships between genes and metabolic or intrinsic characteristics of the patient.

Nutritional Intervention & Gene Expression using Clustering

Within the specific field of nutrigenomics, clustering using data generated by microarray gene expression profiles can be used to identify sub-populations of subjects that respond differently to a given diet intervention. Few works are found using clustering in this field. Papers [Loor et al., 2011] and [Loor and Bionaz, 2012] refer to works with animals. In [Loor and Bionaz, 2012], clustering is used to typify the different kind of diets of a set of cows with more o less fats.

Only two references have been found were clustering is used in a pre-post analysis in nutritional genomics with humans.

The first study is about the effect of different low-energy diets on gene expression [Dahlman et al., 2005]. For this study 40 obese women were randomly assigned to two different dietary interventions of low-energy. In this study, they performed a hierarchical clustering on the genes showing changes in their expression during the dietary intervention to see which of them changed in similar way after diet. Finally, a total of 52 genes were significantly up-regulated and 44 were down-regulated as a result of the intervention, but no diet-specific effect was observed.

The second study [Clement et al., 2004, Viguerie et al., 2005] is an interesting antecedent. It focuses on clustering a human subcutaneous adipose tissue gene expression data set obtained during low-calorie diet intervention to aid in the prediction of 6-month weight loss maintenance. The aims of the study were 1) to identify the best performing clustering method for clustering samples, 2) to identify differential responders to the low calorie diet, and 3) to identify the biological pathways affected during the low-calorie diet by weight maintainers and weight re-gainers. The patterns of expression were compared with that of 17 non-obese subjects. Statistical methods were used to find the relevant set of genes for loose of weight (those of immune or inflammatory response, acute phase response, cellular defense and response to stress). They computed the mean gene expression of every patient for different experimental conditions for microarray and clustered the patients according to the gene expression profiles.

More in detail, they performed an agglomerative hierarchical clustering with average linkage to both mean gene ratios and experiments using the nonparametric statistics. Pearson rank correlation coefficient was used as a measure of similarity. Cluster analysis showed that the pattern of gene expression in obese subjects after 28 day dietary intervention was closer to the profile of non-obese subjects than to the pattern of obese subjects before the intervention.

This work timidly points to a need of *local studies over the different groups of intervention*. They compared the final profiles between intervention (over weight group) or control group. This idea, little exploited yet, is one of those that we collect in our proposal. The idea of decomposing the study in such a way that the local relationships between attributes and the interactions of high order can be better captured.

In fact, one frequent comment in those cited works is that the techniques used till the moment do not allow an easy discovery of the multivariate interactions between several genes which are co-regulated.

Therefore, we think that clustering can bring light to this aspect because it can identify multivariate patterns in a more natural way without requiring clear previous knowledge about the laws governing the target phenomena. Finding a prior substructure of the problem can make easier to get richer models.

2.3 Clustering

Clustering is an approach of unsupervised learning that tries to find hidden structure in unlabeled data. The aim of clustering is to group a set of objects into classes, groups or clusters of similar objects. The definitions of similarity vary from one clustering method to another [Duda et al., 2012].

The general idea behind clustering is finding natural groupings among individuals and thus, the partition of the individuals into clusters or classes.

The different methods of clustering can be grouped depending on several properties. In next list, the clustering approaches are classified by the knowledge representation:

- Structured
  - Hierarchical Clustering
    - Agglomerative (*bottom-up*)
    - Divisive (*top-down*)
  - Bases on Grids or networks such as Self-Organizing Maps
- Unstructured - obtains a single partition of the data instead of a clustering structure such as: k-means, PAM, ISODATA, EM, etc.

Other possible criteria are the following:

- how the clusters are created: agglomerative, divisive or mixing both
- if they need to input a predefined number of clusters
- the model that they are based on: connectivity, centroid, medoids, graphs, density, distribution, etc.
- if they are polythetic: taking all attributes at the same time or those are added one by one
- if they are incremental or not

In this thesis, hierarchical clustering will be used and more detailed description of this family is given below (in Section 2.3.1).

Generally, clustering methods aims to maximize the similarity within the clusters and the dissimilarity between the clusters. One can think that it is desirable that each subject belong clearly to one cluster, but, dealing with real data that is not always possible. For this reason, some clustering approach relaxes the membership concept. The classification of this membership can be the following:

- Strict or exclusive: each subject belong to a unique class
- Hierarchical: each class belongs to a superclass until all subjects are in the same class (root)
- Probabilistic: each subject has a degree of membership to every group
- Subspace: each subject belong to different classes depending on subspace projection.
- Overlapping: the classes are not strictly separated and they can be overlapping other classes.

Currently there are multiple clustering approaches and most of them, have different variants.

2.3.1 Hierarchical Clustering

Advantages of hierarchical clustering include [Berkhin, 2006]: embedded flexibility regarding the level of granularity, ease of handling of any forms of similarity or distance and consequently, applicability to any attribute types. In addition, hierarchical clustering does not need to predefine the number of required clusters.

For our purpose, we use an agglomerative hierarchical clustering with Ward's minimum variance method as agglomeration method (see description in Subsection 2.3.1.1).

Besides, this clustering method is a distance-based method where comparisons between objects are used to guide classes' formation. Since data are heterogeneous, Gower's dissimilarity coefficient is selected, as it handles heterogeneous data (see definition in Section 2.4).

Criteria to cut the dendrogram:

a) Calinski-Harabasz Index (see definition in 2.5). It is used to evaluate the quality of the found clusters. It can be assessed and maximized at every level of the dendrogram.

b) Height of level indexes [15]. The height shows the distance between each level of the dendrogram. Suitable cut of the dendrogram can be derived from big gaps on a visualization.

Calinski-Harabasz Index is used to evaluate the quality of the found clusters. So, at each level of the dendrogram can be assessed. The level index show the distance between each level of the dendrogram. Then, the identification of a suitable cut of the dendrogram can be derived from a visualization of these level indexes.

2.3.1.1 Ward's method

Ward's method [Ward Jr, 1963] is widely used because it is a criterion related to the quantity of information of the clusters, and the produced clusters often are easier to interpret than those obtained with other methods.

Ward's method finds compact and spherical clusters [Clarke et al., 2009], and it is based on merging classes that produce the minimum inter-class variance loss.

2.3.2 Partitional Methods

Partitional clustering methods obtain a single partition of the data instead of a clustering structure, such as the dendrogram produced by a hierarchical clustering. A problem of the partitional methods is the choice of the number of desired clusters which has to be pre-defined.

K-Means: *K-Means* [MacQueen, 1967] is a clustering algorithm belonging to the partitional clustering and that split the data into k clusters. The algorithm is based on reducing squared error between the empirical mean of a cluster (centroid) and the individuals in the cluster [Jain, 2010]. There are different strategies to initialize the k centroids. Once there are k centroids, each individual is classified in the cluster in which the distance to is centroid is the lowest. Then, the centroids are recalculated in base on the individuals that are assigned to each cluster. This process stops when convergence or with a predefined number of iterations.

PAM: PAM [Kaufman and Rousseeuw, 1987] is a partitional clustering algorithm. It is related with k-means and also needs a predefined number of clusters. This algorithm works with a matrix of dissimilarity, where its goal is to minimize the overall dissimilarity between the representatives of each cluster (medoids) and its members. Partitioning of the data into k clusters "around medoids", a more robust version of K-means. Medoids are representative individuals of a cluster whose average dissimilarity to all the individuals in the cluster is minimal. Medoids are similar in concept to means or centroids, but medoids are always members of the data set. The differences with k-means are: the use of medoids instead of centroids and the dissimilarity between subjects is not restricted to be L_2-norm.

2.3.3 *Multiview Clustering*

The *Multiview Clustering* approach is introduced to address the high dimensionality. In our proposed approach, the attribute space is split in different subsets that play different roles in the analysis, named here *thematic blocks*. In principle each *thematic block* is analyzed as an independent view.

The concept of splitting the dimensionality in different groups and then, clustering separately the individuals is found in [Bickel and Scheffer, 2004]. In that work attributes are split into two independent groups according to their meaning.

Other reference using the separately each view is [Yin et al., 2005]. They present a query system where the user can specify a keyword or a small set of keywords for each view, then the system retrieves all attributes in the database relevant for that keyword to perform the clustering. So, the same data can be clustered under different views according to the user specifications.

A different approximation is found in [Li and Shafto, 2011]. The authors present Cross-clustering as a kind of Multiview Clustering. Multiple hierarchical Bayesian clusterings are made; each using a single attribute as separated views, followed by merging the several views in a final single clustering, by merging all trees under certain criteria.

The need of finding consensus techniques to combine the various clusterings obtained under Multiview Clustering is stressed in [Abdullin and Nasraoui, 2012]. Ensemble methods and semi-supervised techniques are proposed among others. This work claims that most of the effort must concentrate on the agreement between the different views or resulting partitions.

In this thesis, *an extension of Multiview Clustering using hierarchical clustering is proposed followed by a process of integrating the results of all views in a single final partition* (the Integrative Multiview Clustering approach, see Section 3.2.1).

2.3.4 Clustering in Nutrition

As said before, clinical studies use traditional basic hypothesis testing or statistical modeling. Clustering is not a very known method in this area. Few works are found in the field using clustering. We refer here a couple of them:

In [Hulshof et al., 1992], dietary attributes and other lifestyle factor are studied. For this task, individuals are grouped using K-means. Since, K-means requires that the number of clusters is specified before the analysis, they run k-means from two to ten classes and then, the selection of the best number of clusters is done based on a screen plot, in which the variance within clusters is represented against the number of clusters.

In [Swaminathan et al., 2012], a two-step cluster analysis revealed two distinct clusters: obesogenic and nonobesogenic. This study clustered the individuals upon diet and physical activity.

Although few works were found, we are convinced that using clustering for finding patterns of response to diet may be of help to assist decision making in diet design.

2.3.5 Clustering of Genetic Data

Most of the effort in this area has been done on the analysis of gene expression. This data usually comes from an interpretation of the Microarray Technology (see Section 2.1.1).

Table 2.2: Classification of the references using clustering in Microarrays

Method	References
Hierarchical clustering	[Eisen et al., 1998, Alizadeh et al., 2000, van't Veer et al., 2002]
Fuzzy clustering	[Xu and Wunsch, 2010, Gasch and Eisen, 2002, Möller-Levet et al., 2003]
Evolutionary clustering	[Ma et al., 2006]
Ant-based clustering	[He and Hui, 2009]
Self-Organizing Maps	[Tamayo et al., 1999, Ghouila et al., 2009]
Biclustering	[Cheng and Church, 2000, Gremalschi and Altun, 2008, Tanay et al., 2002, Prelic' et al., 2006, Ayadi et al., 2012, Yang et al., 2002, Yang et al., 2003, Getz et al., 2000, Kriegel et al., 2009]
Triclustering	[Zhao and Zaki, 2005, Mahanta et al., 2011]

This data is a numeric matrix but several clustering approaches transform it into binary matrix. In Table 2.2, the references are classified by the clustering method that they use.

Clustering has been used to study genes when data can be used for clustering genes or gene-conditions interactions. In the review [Yoo et al., 2012] they state that the results of the clustering of genes can be very valuable asset when those are posteriorly used by gene researchers.

Hierarchical Clustering

In the origins of genetics, simple organisms were analyzed and also, using clustering techniques. In instance, the highly analyzed yeast *Saccharomyces cerevisiae* was used in [Eisen et al., 1998] to cluster gene expression using an Average linkage hierarchical clustering. Genes with similar expression across multiple growth conditions were found using Pearson Correlation similarity.

The similar methodology was used in [Alizadeh et al., 2000], to understand why the treatment of diffuse large B-cell lymphoma is only successful 40% of times. Microarrays of some genes were analyzed: those that have been studied and which may have a relation with this lymphoma or others through normal and malignant lymphocyte samples. Hierarchical clustering was applied to both axes (genes and conditions) using the weighted pair-group method with centroid average and Pearson correlation distance.

In [van't Veer et al., 2002], hierarchical clustering allowed grouping the 98 primary breast tumors on the basis of their similarities measured over the approximately 5,000 significant genes.

Fuzzy Clustering

Fuzzy clustering has been also used in order to establish the relationships between genes and multiple functional categories, and to identify conditional co-regulation of genes [Xu and Wunsch, 2010]. In [Gasch and Eisen, 2002], they use a heuristically modified version of fuzzy k-means clustering to identify overlapping clusters of yeast genes based on published gene-expression data following the response of yeast cells to environmental changes. They conclude that Fuzzy k-means clustering is useful for extracting biological insights from gene-expression data. They found overlapping sets of genes over the condition-specific co-regulation of gene expression.

In [Möller-Levet et al., 2003], a distance for short time-series is presented to compare the changes of different gene expression in different uneven time intervals of the *Saccharomyces cerevisie*. Fuzzy c-means algorithm was used as a template to introduce the metrics for short time-series.

Evolutionary Clustering

In [Ma et al., 2006], an evolutionary clustering algorithm called *EvoCluster* is proposed, handling noise in microarray data. This algorithm was tested with real Microarray data to contrast the effectiveness of this algorithm. Results were compared with previous findings of the same data and viability of EvoCluster is shown.

Ant-based Clustering

In [He and Hui, 2009], an ant-based clustering algorithm (Ant-C) is proposed for gene expression data analysis. Main goal is to cluster the genes based on their expression profile. This algorithm is based on ant colony optimization, a nature inspired algorithm emerging from the collective behavior of social ant colonies. The result of Ant-C is a fully connected network of nodes. Each node represents a gene, and every edge is associated with a certain level of pheromone intensity. Then, minimum spanning tree is applied to break the linkages of the network to generate the resulting clusters. These algorithms were tested with data from yeast, rats and human.

Self-Organizing Maps

In cite [Tamayo et al., 1999], they use GENCLUSTER which is an algorithm based on Self-Organizing Maps for gene-profiling. The program begins with two preprocessing steps so as to improve the ability to detect meaningful patterns. First, the data are passed through a variation filter to eliminate those genes with no significant change across the samples. This step prevents nodes from being attracted to large sets of invariant genes. Second, the expression level of each gene is normalized across the experiment. This focuses attention on the "shape" of expression patterns rather than on absolute levels of expression. And then, SOM is computed.

Multi-SOM consisting of a hierarchy of SOM grids was used to cluster macrophage gene expression data, which aims to reduce the dependency of SOM on the user-specified number of clusters [Ghouila et al., 2009]. A multi level SOM based clustering algorithm was proposed. Through the use of clustering validity indexes, Multi-SOM overcomes the problem of the estimation of clusters number. This approach was tested with macrophage gene expression data generated in vitro from the same individual blood infected with 5 different pathogens. The idea of this approach is the following: first level is used to train data by the SOM in order to decrease the input space complexity. Then the other levels are used to cluster data based on the resulting SOM grid. The output neurons are gradually clustered using multiple SOM grids.

Biclustering

As an alternative to clustering on genomic data, multiple biclustering algorithms have been used. Given a matrix, biclustering aims to cluster rows and columns simultaneously [Cheng and Church, 2000]. These algorithms are based on the idea that *many genes have more than one function in the cellular process and can be co-regulated with more than one cluster of other genes, then it is not rare to find genes that belong to at least two categories [Xu and Wunsch, 2010].*

The main biclustering classes are the following [Madeira and Oliveira, 2004]:

- Biclusters composed by constant values.
- Biclusters with constant values in rows or columns.
- Biclusters coherent values.
- Biclusters with a coherent evolution of their values.

The simpler Biclustering algorithms are those which identify biclusters of constant values. For this reason gene expression matrix are transformed into a binary matrix where the values represented whether a gene is expressed or not. There are many biclustering variants, and inside of them, there are algorithms whose allow overlapping of clusters.

One of the first publications about this type of clustering is [Hartigan, 1972]. In this article it was presented a model for clustering subjects and attributes simultaneously. This model *Direct Clustering* finds clusters searching for values that are equals (constant) in subjects or attributes or both. This type of clustering is currently called "Biclustering".

The algorithm of Cheng and Church [Cheng and Church, 2000] is a reference in biclustering with gene expression data. This algorithm is based on graphs. They use complete bipartite graphs and work adding and deleting nodes in order to find clusters that have low mean squared residue scores (MSR). This algorithm aims to find biclusters with coherent values. As an evolution of this algorithm, in [Gremalschi and Altun, 2008] a Mean Squared Residue Based Biclustering algorithms that can find several clusters whose size is not predefined.

In [Tanay et al., 2002] present a bicluster algorithm for detecting significant biclusters in large expression datasets. Their approach is graph theoretic coupled with statistical modeling of the data.

An important algorithm is Bimax [Prelic' et al., 2006]. This algorithm transform the data into a binary matrix. The model assumes two possible expression levels per gene: no change and change with respect to a specific condition. This algorithm is based on *divide and conquer* strategy.

BicFinder [Ayadi et al., 2012] relies on a new evaluation function called Average Correspondence similarity Index (ACSI) to assess the coherence of a given bicluster and utilizes a directed acyclic graph to construct its biclusters. This algorithm is able to find coherent and overlapping biclusterings.

The FLOC (FLexible Overlapped biClustering) [Yang et al., 2002, Yang et al., 2003] approach generates a set of initial biclusters and iteratively, they are improving their quality. This approach is proposed for solving the problem of *interference noise* identified in the algorithm of Cheng and Church. Interference noise is the noise introduced as random values to fill the missing values of the data.

In [Getz et al., 2000], the authors present the CTWC (Coupled Two-Way Clustering) which establish a generic framework to be used by several clustering techniques. CTWC aims to find submatrix which are significant for the data. They show results using the hierarchical clustering algorithm as SPC (Super-paramagnetic clustering of data) [Blatt et al., 1996] stating that is suitable for microarray data.

In [Kriegel et al., 2009] a survey of clustering techniques for high-dimensional data is introduced. In this survey biclustering is included in *Pattern-Based Clustering*. Also, they describe the *Correlation Clustering* as an improvement of biclustering because this last one cannot include negative or complex correlations. Correlation clustering algorithms assume any cluster being located in an arbitrarily oriented subspace of the data space. Two of the authors of this survey presented a PhD thesis the previous year [Zimek, 2008]. In this thesis, they proposed several algorithms based on correlation.

- The first algorithm is 4C which is based on a density-based clustering paradigm.
- A more robust, more efficient and more effective variant of 4C for flat correlation clustering is COPAC
- HiCO is a hierarchical correlation algorithm. This approach defines the distance between points according to their local correlation dimensionality and subspace orientation, and uses the hierarchical density-based clustering.
- ERiC is also a hierarchical correlation algorithm. This approach derives a local eigen system for a point based on the k-nearest neighbors in an Euclidean space. The resulting set of clusters is hierarchically ordered to provide a hierarchy of subspace clusters.
- CASH is a density-based approaches applying Principal Component Analysis on a local selection of representative points. This approach use the *Hough Transform* which maps points from Euclidean space to a parameter space which is used for assessing the distance between subjects.

Triclustering

It is even possible to find algorithms for finding triclusters so as to extract genes which have similar expression patterns for a set of samples across a set of time points [Zhao and Zaki, 2005, Mahanta et al., 2011].

None of these works model the relationship among gene expression and other physiological attributes.

2.4 Mixed Metrics

There are different types of attributes depending on the values that can have.

- Numerical attributes which are not restricted to a particular values.
- Categorical attributes that contains values belonging to one of several possible categories.
- Binary attributes that can be seen as particular case of categorical attributes with two categories.

Among all attributes, the *active* attributes are those used to construct the clusters and the *illustrative* ones are those who have not been used for the construction of these clusters, but can participate in the interpretation process to enrich the description of the clusters.

Most of the works, referred below on clustering refer to numerical data. However, as soon as clinical information wants to be related with other attribute describing the individuals, qualitative attributes may be also of importance to be included in the analysis. For this purpose, a heterogeneous metric will be used. There are many possibilities in the literature.

Upon [Anderberg, 1973], three main strategies, more extensively discussed in [Gibert and Cortés, 1992], may be followed:

- *Partitioning* the attributes upon their type, then analyzing the dominant type [Lebart et al., 1990];
- *Converting* all attributes to a unique type, conserving as much original information as possible[1]; many authors [Anderberg, 1973, Gibert, 1994], discuss on this line;
- *Compatibility measures* covering any combination of attributes types; the idea is to allow clustering on heterogeneous data matrices without transforming the attribute themselves. Main advantages of this approach are that it is respecting the original nature of data, there is not loss of information, no need to take previous arbitrary

[1] In Statistics, traditionally, symbolic attributes have been converted to a set of binary variables; then, clustering with χ^2 metrics is suitable [Lebart et al., 1990]. In Artificial Intelligence (AI), transformation of quantitative values into a qualitative one is much more popular.

decisions which can bias results, enables study of all types of attributes together, and enables analysis of interactions between attributes of different types. Since distances between individuals are needed in the core of clustering, a function to compute it with heterogeneous data is required. In the literature, several proposals are found.

Some mixed metrics used for clustering real heterogeneous data are described:
Gower [Gower, 1971]. Gower works were the first on the direction of defining similarities, and afterwards distances, for spaces where numerical and qualitative attributes coexists. Briefly, Gower's metrics combines a normalized distance on the absolute values (first order normalized Minkowsky metrics) for the numerical attributes with equality or not for qualitative, taking into account missing data.

The original *Gower's coefficient* definition is formulated as a similarity measure. The symmetric measure is a dissimilarity measure. In [Gower, 1971] Gower demonstrates that the square of the original dissimilarity is a metrics, so it makes suitable for any distance-based method like hierarchical clustering.

In this work the Gower coefficient is used. The Gower's coefficient considers all these types of attributes [D'Orazio, 2012]. Given two objects i, j, the Gower's coefficient is computed as in next Equation 2.1.

$$\begin{aligned}
d_G(i,j) &= \frac{\sum_k \delta_{ijk} * dist_{ijk}}{\sum_k \delta_{ijk}} \\
dist_{ijk} &= \begin{cases} 0 & \text{if attribute}_k \text{ is binary} \wedge x_{ik} = x_{jk} = TRUE \\ 0 & \text{if attribute}_k \text{ is categorical} \wedge x_{ik} = x_{jk} \\ \dfrac{|x_{ik} - x_{jk}|}{max_k - min_k} & \text{if attribute}_k \text{ is numerical} \\ \dfrac{|r_{ik} - r_{jk}|}{max(r_k) - 1} & \text{if attribute}_k \text{ is ordered} \\ (r_{ik} = category\ index) & \\ 1 & \text{otherwise} \end{cases} \\
\delta_{ijk} &= \begin{cases} 0 & \text{if } x_{ik} = MISSING \text{ or } x_{jk} = MISSING \\ 0 & \text{if the attribute is asymmetric binary and } x_{ik} = x_{jk} = 0 \\ & \text{or } x_{ik} = x_{jk} = FALSE \\ 1 & \text{otherwise} \end{cases}
\end{aligned} \tag{2.1}$$

Gowda & Diday [Gowda and Diday, 1992]. It is built with three components called *position, span and content*: $D_k(i,i') = D_p(i,i') + D_s(i,i') + D_c(i,i')$. In fact, D_p is the same component for numerical attributes as Gower, and for qualitative ones it is null, while D_s, D_c are considering the number of objects in each modality as well as their intersections. Also suitable for other types of data, not considered in this research.

Gibert 91. Introduced in [Gibert and Cortés, 1992,Gibert and Cortés, 1997]: $d^2_{(\alpha,\beta)}(i,i') = \alpha d^2_{\zeta}(i,i') + \beta d^2_{\mathbb{Q}}(i,i')$, being $(d^2_{\zeta}(i,i'))$ the normalized euclidean metrics for numerical attributes, and $(d^2_{\mathbb{Q}}(i,i'))$ the χ^2 metrics[1] for qualitative. Actually, it is a family of metrics indexed on $(\alpha,\beta) \in [0,1]^2$; there is a proposal for weighting on the basis of dimensionality of both spaces and on a normalizing factor [Gibert and Sonicki, 1999b]. Successfully applied to several real *IDSS* [Gibert and Annicchiarico, 2003,Gibert and Roda, 2000,Gibert and Cortés, 1998,Gibert and Sonicki, 1999b,Cheeseman and Oldford, 2012].

Ichino & Yaguchi [Ichino and Yaguchi, 1994]. A generalization of Minkowsky metrics based on a new formal model supporting a single expression for all attributes: $\Phi(x_{ik}, x_{i'k}) = |x_{ik} \oplus x_{i'k}| - |x_{ik} \otimes x_{i'k}| + \gamma(2|x_{ik} \otimes x_{i'k}| - |x_{ik}| - |x_{i'k}|)$ on the basis of the operators $\oplus, \otimes$ defined in the model, $\gamma \in [0, 0.5]$. For numerical attributes, definition coincides with Minkowsky metrics; for qualitative ones, it only distinguishes equality or not.

L'Eixample L'Eixample [Sànchez-Marrè et al., 1998] [Sànchez-Marrè et al., 1999] is a normalized weight-sensitive distance function.

For the most important numerical attributes - that is, weight $> \alpha$ - the distance is computed based on their qualitative values (a previous discretization is done for numerical attributes). This implies that relevant attributes having the same qualitative value are equals, and having different qualitative values are very different, even when a continuous measure would be very small. And for those less relevant ones the distance is computed based on their quantitative values. This metric consider numerical and categorical attributes and distinguish between ordered and non-ordered categorical attributes.

Ralambondrainy [Ralambondrainy, 1995]. As a matter of fact, it is defined in the same way as Gibert's metrics, but the proposed coefficients for weighing the two components of the metrics are calculated on the basis of previous works of the same author [Ralambondrainy, 1988], using a much more formal paradigm, taking into account the inertia of the groups or the norm of operators.

2.5 Cluster Validity

Cluster validity methods aim at the quantitative evaluation of the results of the clustering algorithms. These methods try to cope questions such as "how many clusters are in the dataset?" or "is there a better partitioning for our dataset?" [Halkidi et al., 2001].

Most of the works done in cluster validation on unsupervised context are centered in the internal validation of the clusters [Arbelaitz et al., 2013,Dimitriadou et al., 2002,Maulik and Bandyopadhyay, 2002,Milligan and Cooper, 1985,Dubes, 1987,Halkidi et al., 2001,Halkidi

[1] This means that difference between qualitative values refers the quantity of information involved [Benzécri, 1973]

et al., 2002a, Halkidi et al., 2002b, Bezdek and Pal, 1998] using what is known as a *cluster validity index (CVI).*

Previous works have shown that there is no single CVI outperforming the rest [Brun et al., 2007, Dimitriadou et al., 2002, Maulik and Bandyopadhyay, 2002, Milligan and Cooper, 1985] and there are few works that compare several cluster validity indexes in order to draw some general conclusions [Arbelaitz et al., 2013, Brun et al., 2007, Dimitriadou et al., 2002, Milligan and Cooper, 1985] and no general guidelines exists to help the analyst to choose the best CVI in front of a real case.

A reference in this area is [Milligan and Cooper, 1985] which compares 30 cluster validity index using artificial datasets and Monte Carlo simulation.

Other popular study is [Dubes, 1987] which compares Davies-Bouldin Index and a modification of Hubert Γ Statistic using Monte Carlo method.

In [Bezdek and Pal, 1998], they evaluate two clustering methods (hard c-means and simple linkage) and the results are compared using Davies-Bouldin, Hubert's statistics, Dunn and variants of Dunn indexes. They suggest in this study that there is not an index which provides consistent results across different clustering algorithms and data structures.

In [Dimitriadou et al., 2002], the performance of 15 indexes for determining the number of clusters in 162 synthetic binary datasets is analyzed. Based on the ability to recommend the correct number of clusters, they proposed to use Ratkowsky-Lance and Davies-Bouldin followed by Calinski-Harabasz and Xun indexes regarding to point the correct number of classes.

A comparison of a new proposed cluster validity index against others is performed in [Maulik and Bandyopadhyay, 2002]. In this work they conclude that there is not a unique index which it is good enough to determine the number of clusters.

In [Brun et al., 2007], there is a comparison of different cluster indexes from different types, some evaluating the properties of the partition itself (internal criteria), some comparing with a reference partition (external criteria that usually corresponds to accuracy error measurement) and some comparing several partitions among them (relative criteria). Authors claim that the external criteria are better when a reference partition for comparing is available. In other cases, they conclude that Silhouettes index outperforms in their experiments.

A comparison of the most popular cluster validity indexes are performed in [Arbelaitz et al., 2013]. In this work the CVIs are compared using 720 synthetic datasets and 20 datasets from the UCI. The synthetic datasets are all possible combinations of the following 5 factors: number of clusters (2,4,8), dimensionality (200,400,800), cluster overlap (yes, no), cluster density (equal, asymmetry) and noise level (without, with). For each dataset 3 clustering methods (k-means, Ward and Average-linkage) are run using k from 2 to , being n the size of the dataset. Both synthetic and UCI datasets are available online along with their results. This work concludes that indexes with better performance seem to be Silhouettes, David-Bouldin and Calinski-Harabasz.

Those indexes evaluate the structure of the resulting clusters, whether the clusters are more separated between them or whether the cluster are more cohesive. Nevertheless, they can not evaluate the meaning of these clusters.

Thus, structural validity does not necessarily ensure the *usefulness* of clustering. *Meaningfulness* is, also, a key feature for guaranteeing decision-making support. Usefulness and understandability are part of the characteristics required in a data mining solution according to the seminal paper [Fayyad et al., 1996].

The most commonly used internal Cluster Validity Indexes (CVIs) and indicators in the literature are described in this section. As it was mentioned, they are commonly used in cluster validation because they do not need additional information other than data and clusters themselves. The internal validation evaluates the resulting clusters in base to its topography or structure. This evaluation is mostly based on the compactness (cohesion) of the clusters and the separation between clusters. Literature on indexes to measure this internal validity is abundant [Arbelaitz et al., 2013, Brun et al., 2007, Dimitriadou et al., 2002, Maulik and Bandyopadhyay, 2002, Milligan and Cooper, 1985, Dubes, 1987, Halkidi et al., 2001, Halkidi et al., 2002a, Halkidi et al., 2002b, Bezdek and Pal, 1998, Gordon, 1999, Kim and Ramakrishna, 2005, Meilă, 2007, Hennig, 2013b]. Most of the indexes estimate the cluster cohesion (within or intra-variance), the cluster separation (between or inter-variance) or combine both to compute a quality measure [Kim and Ramakrishna, 2005].

Table 2.3 contains the list of the indexes and indicators with the range of values they can have and its optimal value.

Table 2.3: Cluster Validity Indexes

Index	Range	Optimal Value	Index	Range	Optimal Value
Entropy	$[0,)$	Minimum	Dunn-like	$[0,)$	Maximum
Diameter (Δ)	$[0,)$	Minimum	CH	$[0,)$	Maximum
WG	$[0,)$	Minimum	$\hat{\Gamma}$	$[-1, 1]$	Maximum
$\overline{W}$	$[0,)$	Minimum	Silhouettes	$[-1, 1]$	Maximum
WSS	$[0,)$	Minimum	BH	$[-1, 1]$	Maximum
$\overline{B}$	$[0,)$	Maximum	WBR	$[0,)$	Minimum
Separation (δ)	$(0,)$	Maximum	C-Index	$[0, 1]$	Minimum
Sindex	$(0,)$	Maximum	DB	$[0,)$	Minimum
Dunn	$[0,)$	Maximum			

The following formulation of all indexes is provided under a common notation. Given a dataset $\mathcal{X}$ and a partition P.

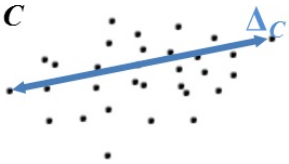
$\mathcal{X}$ is composed by n individuals $I = \{i_1, \dots, i_n\}$ and K attributes $X = \{X_1, \dots, X_K\}$. The partition $P = \{C_1, \dots, C_\xi\}$ contains ξ clusters where $n_c = card(C)$, $C \in P$ and $C \bigcap C' = \oslash$, $C, C' \in P$. Being $d(i, i') = \sqrt{\sum_{k=1}^{K}(x_{i_k} - x_{i'_k})^2}$ for all i, $i' \in I$, the 17 CVIs are defined in the following way.

Entropy: The entropy index measures the entropy associated with partition P [Meilă, 2007]. Entropy is always non-negative. It takes value 0 only when there is no uncertainty, namely when there is only one cluster. Entropy is measured in bits. The uncertainty of 1 bit corresponds to a clustering with $\xi = 2$ and $P(1) = P(2) = 0.5$. Note that the uncertainty does not depend on the number of objects in I but on the relative proportions of the clusters. This index is defined as following:

$$Entropy = -\sum_{C \in P} \frac{n_c}{n} log\left(\frac{n_c}{n}\right) \tag{2.2}$$

Maximum Cluster Diameter: The maximum cluster diameter (Δ) is the maximum distance between any two points that belongs to the same cluster [Hennig and Liao, 2010]. In other words, it is defined by the higher diameter among all the clusters belonging to P (see Figure 2.3).

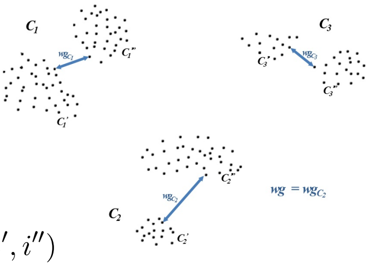
Figure 2.3: Maximum Cluster Diameter



$$\Delta = \max_{C \in P} \Delta_c, \quad \Delta_c = \max_{i,i' \in C} d(i,i') \tag{2.3}$$

Widest gap: The widest gap index (wg) is the maximum within-cluster gap for all clusters. The widest within-cluster gap (wg_c) is defined as the largest link in within-cluster minimum spanning tree [Hennig, 2013b] (see Figure 2.4).

Figure 2.4: Widest gap



$$wg = \max_{C \in P} wg_c, \tag{2.4}$$

$$wg_c = \max_{\substack{C', C'' \\ C' \bigcap C'' = \oslash \\ C' \bigcup C'' = C}} g(C', C''), \qquad g(C', C'') = \min_{\substack{i' \in C',\ i'' \in C'' \\ C' \bigcap C'' = \oslash \\ C' \bigcup C'' = C}} d(i', i'')$$

Average Within-Cluster Distance: The average within-cluster distance ($\overline{W}$) is the average of the distances between all pairs of objects within the same cluster.

$$\overline{W} = \frac{\sum_{C \in P} W_c}{\sum_{C \in P} n_c(n_c - 1)}, \quad W_c = \sum_{i,i' \in C} d(i,i') \tag{2.5}$$

Within Cluster Sum of squares: The Within Cluster Sum of squares (WSS) is the sum of squared distances between all pairs of objects within a cluster [Halkidi et al., 2002b]. Where i_c = barycenter of cluster C

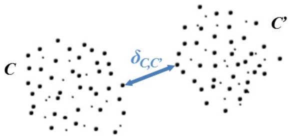
$$WSS = \sum_{C \in P} WSS_c, \quad WSS_c = \sum_{i \in C} d(i, i_c)^2 \tag{2.6}$$

Average Between-Cluster Distance: The average between-cluster distance ($\overline{B}$) is the average of all distances between pairs of objects which do not belong to the same cluster. Given a pair of classes $C, C' \in P$

$$\overline{B} = \frac{\sum_{\substack{C,C' \in P, \\ C \neq C'}} dist(C, C')}{\sum_{\substack{C,C' \in P, \\ C \neq C'}} n_c n_{c'}}, \quad dist(C, C') = \sum_{i \in C} \sum_{i' \in C'} d(i, i') \tag{2.7}$$

Minimum Cluster Separation The minimum separation index (δ) is the minimum distance between any two objects that do not belong to the same cluster. In other words, it is defined by the lower separation among all the clusters (See Figure 2.5). $\delta_{c,c'}$ is highly related with wg_c measure. $\delta_{c,c'}$ finds gaps between clusters whereas wg_c finds gaps inside each cluster.

Figure 2.5: Minimum Cluster Separatation



$$\delta = \min_{C,C' \in P} \delta_{c,c'}, \quad \delta_{c,c'} = \min_{i \in C, i' \in C'} d(i, i') \tag{2.8}$$

Separation Index: The Separation index ($Sindex$) is based on the distances for every point to the closest point not in the same cluster. The separation index is the mean of the S smallest separations [Hennig and Liao, 2010,Hennig, 2013b] being S a certain proportion of the dataset. This allows formalizing separation less sensitive to a single or a few ambiguous points.

$s \in [0, 1], S = E[ns]$ an object $i \in I, c(i)$ is the class of object i
$\forall i \in I, sep(i) = \min_{i' \notin c(i)} d(i, i')$
Being $\{sep(i)_m\}_{m=1:n}$ the sorted sequence where $sep(i)_m \leq sep(i)_{m+1}$

$$Sindex = \frac{\sum_{m=1}^{S} \{sep(i)_m\}}{S} \tag{2.9}$$

Dunn: Dunn Index (D) is a cluster validity index for crisp clustering proposed in Dunn (1974) [Dunn, 1974]. It attempts to identify "compact and well separated clusters" [Halkidi and Vazirgiannis, 2002]. If a data set contains well-separated clusters, the distances among the clusters are usually larger than the diameters of the clusters. We present a formulation from [Dimitriadou et al., 2002]: (see Section 2.5 for $\delta_{c,c'}$ and 2.5 for Δ_c)

$$D = \frac{\min_{C,C' \in P} \delta_{c,c'}}{\max_{C \in P} \Delta_c} \tag{2.10}$$

Dunn-like: Dunn-like index is one of the generalizations of Dunn index [Dunn, 1974] proposed by Bezdek and Pal [Bezdek and Pal, 1998]. It attempts to identify compact and well separated clusters. If a data set contains well-separated clusters, the distances among the clusters are usually larger and the diameters of the clusters are expected to be smaller. From all generalizations proposed in [Bezdek and Pal, 1998,Pal and Biswas, 1997], the one available in FPC-R package [Hennig, 2013a] is the one substituting point to point distance of the original Dunn index by average inter-class distance in numerator and average point to point intra-class distance in the denominator. This version is more robust than the original Dunn Index [Halkidi et al., 2001,Bezdek and Pal, 1998].

$$D = \frac{\min_{C,C' \in P} \overline{\delta_{c,c'}}}{\max_{C \in P} \overline{\Delta_c}}, \quad \overline{\delta_{c,c'}} = \frac{\sum_{i \in C, i' \in C'} d(i,i')}{n_c n_{c'}} \qquad \overline{\Delta_c} = \frac{\sum_{i,i' \in C} d(i,i')}{n_c(n_c - 1)} \tag{2.11}$$

Calinksi-Harabasz Index: Calinski-Harabasz index [Caliński and Harabasz, 1974] is based on getting a compromise between both the between-cluster distances and the within-cluster distances. The Calinski-Harabasz index (CH) is defined as following:

$$CH = \frac{BSS/(\xi - 1)}{WSS/(n - \xi)}, \quad BSS = \sum_{C \in P} n_c d(i_c, \bar{i})^2 \tag{2.12}$$

i_c is the barycenter of the cluster C $\bar{i}$ is the barycenter of I, WSS is already defined. This index is interpreted as higher values as better clustering partition.

$\hat{\Gamma}$ (Normalized Hubert Gamma Coefficient): This index is a Pearson version of Hubert's gamma coefficient [Halkidi et al., 2001]. It gives information on how good the clustering is as an approximation of the dissimilarity matrix. It introduces an auxiliary indicator Y such that evaluates to 1 for pairs of objects in the same cluster and to 0 otherwise. Being D the matrix with distances between subjects; the $\hat{\Gamma}$ index is defined as the correlation between D and Y:

$$y(i,i') = \begin{cases} 1 & c(i) = c(i') \\ 0 & c(i) \neq c(i') \end{cases}, \quad \hat{\Gamma} = \frac{\sum_{i,i' \in I}(d(i,i') - \overline{d})(y(i,i') - \overline{y})}{\sum_{i,i' \in I}(d(i,i') - \overline{d})^2 \sum_{i,i' \in I}(y(i,i') - \overline{y})^2} \tag{2.13}$$

Being $c(i)$ the class of i, $\overline{d}$ the barycenter of all distances and $\overline{y}$ the barycenter of Y.

It is especially useful when clustering is used for dimensionality reduction. This index takes values between -1 and 1.

Silhouettes: The Silhouettes Index [Rousseeuw, 1987] provides a succinct graphical representation of how well each object lies within its cluster.

In principle, this index is assessed for each object, but in order to be used is reduced to the average of all dataset or the average for each cluster.

For each object $i \in I$, let

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{2.14}$$

$a(i)$: average dissimilarity of i with all other data within the same cluster. $a(i)$ shows us how well clustered i is to the cluster assigned (smaller value means better matching).

$$a(i) = \frac{\sum_{i,i' \in C} d(i, i')}{n_c - 1}$$

$b(i)$: the lowest average dissimilarity of i with the data of another single cluster. The cluster with this lowest average dissimilarity is said to be the "neighbouring cluster" of i as it is, aside from the cluster i is assigned, the cluster in which i fits best.

$$b(i) = \min_{C' \neq C} \frac{\sum_{i \in C, i' \in C'} d(i, i')}{n_c - 1}$$

Then, the value of $s(i)$ belongs to $[-1, 1]$ and a higher value indicates that i is better clustered.

The Silhouettes index of a cluster C is

$$S_C = \frac{\sum_{i \in C} s(i)}{n_c}$$

The overall Silhouettes index for a given dataset with a partition is

$$Silhouettes = \frac{\sum_{i \in I} s(i)}{n} = \frac{\sum_{C \in P} n_c S_C}{\sum_{C \in P} n_c} \tag{2.15}$$

Baker and Hubert Index (BH): The Baker and Hubert Index (BH) [Baker and Hubert, 1975] is a variant of Goodman and Kruskal's Gamma [Goodman and Kruskal, 1954]. Comparisons are made between all within-cluster distances and all between-cluster distances. A comparison is considered to be *concordant* if a within-cluster distance is strictly less than a between-cluster distance.

$$BH = \frac{S^+ - S^-}{S^+ + S^-} \tag{2.16}$$

S^+: number of concordant quadruples,

S^- : number of discordant quadruples.

For this index, all possible quadruples (q, r, s, t) of input parameters are considered.

A quadruple (q, r, s, t) is called *concordant* if one of the following two conditions is true:

- $d(q, r) < d(s, t)$, q and r are in the same cluster, and s and t are in different clusters.

- $d(q,r) > d(s,t)$, q and r are in different clusters, and s and t are in the same cluster.

By contrast, a quadruple is called *discordant* if one of following two conditions is true:

- $d(q,r) < d(s,t)$, q and r are in different clusters, and s and t are in the same cluster.
- $d(q,r) > d(s,t)$, q and r are in the same cluster, and s and t are in different clusters.

Within Between Ratio: The Within Between Ratio (WBR) Index is the ratio between the average within-cluster distance and the average between-cluster distance. Then, as lower is the value means that the partition P is more compact and more separated. See the previous definitions of $\overline{W}$ and $\overline{B}$

$$WBR = \frac{\overline{W}}{\overline{B}} \tag{2.17}$$

C-Index: The C-index (C) [Hubert and Levin, 1976, Gordon, 1999] is computed using the within-cluster distances.

$$C = \frac{W - W_{min}}{W_{max} + W_{min}} \tag{2.18}$$

Where, W is Sum of distances over all pairs of objects from the same cluster.

$$W = \sum_{C \in P} \sum_{i,i' \in C} d(i,i')$$

Being $\{d_m\}_{m=1:n^2}, d_m \leq d_{m+1}$ the ordered list of distances between all possible pairs of objects.

$$W_{min} = \sum_{m=1}^{n_W} d_m, \quad W_{max} = \sum_{m=n^2-n_W+1}^{n^2} d_m, \qquad n_W = \sum_{C \in P} n_c^2$$

Let n_W be the number of those pairs. Then W_{min} is the sum of the n_W smallest distances if all pairs of objects are considered (the objects can belong or not to the same cluster). Similarly, W_{max} is the sum of the n_W largest distances out of all pairs.

Davies-Bouldin: The Davies-Bouldin Index [Davies and Bouldin, 1979] is a cluster separation measure. The overall index is defined as the average of indexes computed from each individual cluster. An individual cluster index is taken as the maximum pairwise comparison involving the cluster and the other clusters in the solution.

$$DB = \frac{1}{\xi} \sum_{C \in P} max_{C' \in P, C' \neq C} \left(\frac{S_{p_c} + S_{p_{c'}}}{d_p(C,C')} \right), \tag{2.19}$$

$$d_p(C,C') = \sqrt[p]{\sum_{k=1}^{K} |\overline{X_{c_k}} - \overline{X_{c'_k}}|^p},$$

$$S_{p_c} = \sqrt[p]{\frac{\sum_{i \in C} d_p(i,i_c)^p}{n_c}}$$

Where, i_c is the barycenter of the cluster C and it is defined as $i_c = (\overline{x_{c_1}}, \ldots, \overline{x_{c_k}}), \overline{x_{c_k}} = \frac{\sum_{i \in C} x_i}{n_c}$;

p is the Minkowski factor ($p = 1$ Manhattan distance, $p = 2$ Euclidean distance)

S_{p_c} is the dispersion measure of a cluster C (for $p = 1$ the average distance of objects in cluster C to the barycenter of cluster C; for $p = 2$ the standard deviation of the distance of objects in cluster C to the barycenter of cluster C).

2.6 Cluster Interpretation

In many data mining processes there is still a gap between raw data mining results and effective decision-making. Post-processing data mining results to approach them to the decision makers is crucial for an impact of the analysis on reality. Some works point to this issue [Cortez and Embrechts, 2012, Gibert et al., 2013] and for the particular case of clustering interpretation and conceptualization of clusters is a key issue to this purpose. Some ideas have been developed in previous works [Barnard et al., 2001, Gibert et al., 2008a, Gibert, 2014, Gibert et al., 2012, Siponen et al., 2001].

Cluster interpretation is part of the post-processing step in the data mining process. The interpretation of clusters is an important function when presenting the results to experts. In the literature, the characterization or interpretation of the classes is also termed *Cluster Profiling* [Sarstedt and Mooi, 2011].

The main assumption is that the resulting clusters are different, at least, in base to the data used to create them. Thus, one could expect that not every attribute will be of equal importance when describing the different groups (clusters). Therefore, a comparative analysis between the attributes that have similar values together in the different clusters can be conducted, which provides new insights into the complexity of the cluster description.

Cluster interpretation is usually made by examining the cluster centroids that are built as the average of attributes inside each cluster. In [Sarstedt and Mooi, 2011] it is stated that clusters are distinguishable only if certain attributes exhibit significantly different means in some clusters, at least from a data perspective. This significance is often assessed by comparing the clusters with independent t-test samples or ANOVA. In [Cecere and Abreu, 2010], the attributes used in the clustering are ranked using the *logWorth* index, based on p-values which are usually assessed with the χ^2-Independence Test.

In [Geurts et al., 2003], the resulting two clusters are described by association rules. In this work, they limit the resulting rules to have at least 30% of support. Then, the obtained frequent itemsets are ranked by interestingness measure.

Some efforts also rely on visualizing the attributes through the resulting clusters. In [Haughton et al., 2009], the cluster averages of the standardized attributes are displayed in a parallel plot. In [Talavera and Gaudioso, 2004], the conditional probability of the attributes against the cluster are represented for the categorical attributes. In [Gibert et al., 2008a], the *Class Panel Graphs* are introduced as a graphical representation in the form of panels displaying the conditional distributions of the attributes against the classes (see Section 2.7).

Principal Component Analysis (PCA) is one of the most known techniques in the field of multivariate analysis (see Section 2.8 for a introduction of a classical approach of PCA). A common practice in PCA is to interpret the principal components for eliciting latent attributes that are implicitly measured in the data set and associated with factorial components. Contribution of original attribute to the principal components is used for this interpretation and this is measured by using the angles of the attributes projected over the factorial space, which are directly linked to the correlation between attributes and principal components. Additionally, [Lebart et al., 2000] introduce a statistical test (*Test-Value*) that provides objective criteria to identify the main contributing attributes to a certain principal component. This test is based on the comparison of means/percentages within the classes with respect to the global sample indicator. Although Test-Value comes from the multivariate analysis field, it can help in the interpretation of clusters - as will be shown in this thesis. In Section 6.3, the significances of the attributes obtained with the Kruskal-Wallis test, or the χ^2-Independence test, were compared according to the type of attribute with those obtained using *Test-Value*, and it is seen that for the clustering context, *Test-Value* subsumes the results of Kruskal-Wallis and the χ^2-Independence tests, and *Test-Value* provides more sensitivity (identifying more characteristic attributes than the classical tests). We have not found much in the literature providing interpretation methodologies based on significance tests.

Descriptive statistics and statistical tests are useful to compare an attribute through the different cases, but these methods have the drawback that relates only one attribute with the different clusters. Frequent itemsets seem to be suitable for finding values of different attributes that occurs in every cluster. The drawback of the algorithm for finding frequent itemsets is that only handles categorical or binary data, and thus, it is a problem to find a good discretization for the numerical attributes.

In our work, we will focus on research about the use of statistical tests (see following Subsection 2.9) for cluster interpretation. Also, Class Panels Graphs will be used in the cluster interpretation step of our work. Frequent itemsets are explored in Section 7.2.

In this thesis, we focus on research about a methodology that allows to characterize the clusters. In this methodology, the original *Test-Value* (see definition in Section 2.9.3) is introduced and generalized to reach our goal.

2.7 Class Panel Graphs

The Class Panel Graphs [Gibert et al., 2008a] is a graphical representation in form of panel, containing attributes in the columns and classes in the rows. Conditional distributions of the attributes through the classes are shown per column. They can be represented either via multiple histograms or via multiple boxplots [Gibert et al., 2005]. It shows in a single panel the information of a number of variables through the classes. In a very compact way it provides perspective of the variables behaving particularly for a certain class and supports the interpretation process.

This visualization is interesting for identifying the specific behavior of attributes in a certain class and thus, for better understanding the meaning of the classes [Gibert et al., 2010a]. *Class Panel Graphs* has been successfully used in previous applications [Gibert and Sonicki, 1999a, Gibert et al., 2010a, Gibert et al., 2010b] to present results to experts in support of a class-conceptualization process, and to assess the profiles associated with the classes.

Currently, there is not any available software in the literature that allow creating this kind of representation. In Chapter 4.2.4 there is described our own implementation.

2.8 Principal Component Analysis

Although Principal component analysis (PCA) [Pearson, 1901] appears at the beginning of the XXth century it becomes popular in the late 50's when computer had sufficient capacity. PCA is a multivariate statistical method that finds a reduced set of factors keeping as much information as possible from the original dataset. The factors are orthogonal and they are linear combinations of original attributes. The quantity of information from the original dataset conserved in each factor coincides with the eigenvalues of the covariance matrix of the original dataset. The factors are defined by the eigenvectors. From an algebraic point of view, vectorial base changing operations are found by means of diagonalization of the covariance matrix build over X, to get the most informative projection of the original dataset. From a geometrical point of view, the most informative orthonormal rotation of the original attributes is found.

Given the matrix $X = \{X_1, \ldots, X_K\}$ of K attributes and n objects and being W a diagonal matrix of individual weights, the matrix $Cov(X) = X^T W X$ is diagonalized

$$X^T W X \overline{u} = \lambda \overline{u}$$

The solutions provide the eigenvectors $u_\alpha, \alpha = \{1, .., K\}$ and corresponding eigenvalues $\lambda_\alpha, \alpha = \{1, .., K\}$. The Principal Components are linear combinations of original attributes and $P_\alpha = \sum_{\alpha=1}^{K} u_\alpha X\_alpha$.

Sorting both u_α, λ_α according to decreasing λ_α the first r Principal Components such that $\sum_{\alpha=1}^{r} \lambda_\alpha \geq 0.80$ are conserved.

However, in most of real applications the first factorial plane is analyzed, the one determined by $\langle \overline{u}_1, \overline{u}_2 \rangle$, as it is the one conserving as much information as possible from the original dataset [Benzécri, 1973]. From the 90's on relevant contributions in this field at the level of interpreting results tools [Lebart et al., 2000] like *Test-Value* as one of the contributions of this thesis.

In this thesis PCA has been used to analyze synergies and oppositions of CVI and *Test-Value* had been transferred to a clustering context.

2.9 Statistical Tests

Statistical tests are introduced by Fisher in the 20's as objective procedures to assess whether sample differences with regards to assessed theoretical probabilistic models are relevant or not. In this thesis some of them are used for different purposes. In all cases the evaluation of the test-statistics over sample data give a *p-value* and the decision rule is always the same: $if\ p - value < \alpha\ then\ H_o\ is\ rejected$ (under risk α, α usually $0.05, 0.01, 0.1$). Equivalently, tests can be solved by *critical points.*

2.9.1 Kruskal-Wallis Test

Kruskal-Wallis test [Kruskal and Wallis, 1952] is for numerical attributes. It is a non-parametric method for testing whether several samples come from the same distribution or not. When groups are induced by a qualitative attribute it is equivalent ti see if the numerical and qualitative attribute are independent. When the Kruskal-Wallis test leads to significant results, then at least one of the samples comes from different population of the other samples. As usual, the test does not identify where the differences occur or how many differences actually occur. Since it is a non-parametric method, the Kruskal-Wallis test does not imply distributional assumptions, unlike the analogous one-way analysis of variance (ANOVA).

$$\begin{array}{l} H_o : f_{X|C} = f_X\ \forall C \in P \\ H_1 : \exists C \in P\ f_{X|C \neq f_X} \end{array} \qquad (n-1)\frac{\sum_{c=1}^{P} n_c(\bar{r}_{c\cdot} - \bar{r})^2}{\sum_{c=1}^{P}\sum_{j=1}^{n_c}(r_{cj} - \bar{r})^2} \sim \chi^2_{n-1} \tag{2.20}$$

where, n is the total number of observations across all groups, P is the number of groups, n_c is the number of observations in group c, r_{cj} is the rank (among all observations) of observation j from group c. $\bar{r}_{c\cdot} = \frac{\sum_{j=1}^{n_c} r_{cj}}{n_c}$ is the mean of the ranks assigned to a group c and $\bar{r} = \frac{1}{2}(n+1)$ is the average of all the r_{cj}.

Rejecting H_o means that the numerical and qualitative attributes have some association.

2.9.2 χ^2-Independence Test

χ^2-independence test (Pearson's χ^2 test) [Plackett, 1983] is a non-parametric method for testing whether two qualitative attributes are independent (null hypothesis), or not (alternate hypothesis). For this, the following statistics is used

$$\begin{array}{l} H_o : X, Y \text{ are independent} \\ H_1 : X, Y \text{ are associated} \end{array} \qquad \sum_{x=1}^{C_X}\sum_{y=1}^{C_Y} \frac{(n_{xy} - \frac{n_x n_y}{n})^2}{\frac{n_x n_y}{n}} \sim \chi^2_{(C_X-1)(C_Y-1)} \qquad (2.21)$$

where, C_X is the number of categories of the attribute X, C_Y is the number of categories of the attribute Y, n is the total number of subjects, n_x is the number of subjects with the xth category of X, n_y is the number of subjects with the yth category of Y and n_{xy} is the observed number of subjects with the xth category of X and the yth category of Y.

Rejecting the null hypothesis means that the two attributes (X and Y) have some association.

2.9.3 *Test-Value*

The *Test-Value* was introduced by Lebart in the context of identifying the most important attributes in clusters found over a factorial plane [Lebart et al., 2000].

The *Test-Value* is used to select the most relevant attributes for each class. This test helps the interpretation of the classes based on comparisons of means or percentages within classes with averages or percentages of all elements in the sample, depending if the attribute is numerical or qualitative.

To characterize a class with numeric attributes, the mean of each attribute within the class is compared with the global average of the attribute. The *Test-Value* is computed by the following statistics:

$$\begin{array}{l} H_o : \mu = \mu_C \\ H_1 : \mu \neq \mu_C \end{array} \qquad \frac{\bar{X}^c - \bar{X}}{\sqrt{(1 - \frac{n_c}{n})\frac{s^2}{n_c}}} \sim t_{n_c - 1} \qquad (2.22)$$

where, $\bar{X}$ is the sample mean of attribute X, $\bar{X}^c$ is the sample mean of X within the class C, n is the number of objects in the total sample, n_c is the number of objects in class c and s is the standard deviation of X.

Rejecting the H_o means that the attribute X is significant for class C.

Similarly, a category of a qualitative attribute characterize a class if their abundance in the class is considered significantly higher than would be expected for their presence in the global sample. This criterion is mainly used to highlight the category which characterizes the better the class.

Given a qualitative attribute X, one of its categories s and a particular class C. The *Test-Value* can be defined for qualitative attributes as following:

$$\begin{array}{l} H_o : \pi_s = \pi_s | c \\ H_1 : \pi_s \neq \pi_s | c \end{array} \qquad \frac{\frac{n_{sc}}{n_c} - \frac{n_s}{n}}{\sqrt{(1 - \frac{n_c}{n})\frac{\frac{n_s}{n}(1 - \frac{n_s}{n})}{n_c}}} \sim Z \qquad (2.23)$$

where, n is the number of objects in the sample, n_s is the number of objects in the sample with the category s, n_c is the number of objects of class c and n_{sc} is the number of objects in class c with the category s.

Rejecting the H_o means that the category s of attribute x is significant for class C.

These tests include the correction for the variance estimates in finite populations in the denominator, as the comparison is performed with terms that refer to nested samples (i.e $\overline{X}^c$ is computed with a subset of elements from those used to compute $\overline{X}$).

In our proposal, they are used and adapted for a new methodology to interpret the clusters in an automatic way.

2.9.3.1 Properties

- *Test-Value* subsumes Kruskal-Wallis test and χ^2-Independence test.

  If Kruskal-Wallis test/χ^2-Independence test is significant then *Test-Value* is significant. See details in Section 6.3.

- *Test-Value* is sensitive to the class size(n_c): when the class size decrease, the statistic value also decreases. Therefore, the same mean or proportion lose significance with smaller classes. See details in Section 3.3.

2.10 Intelligent Decision Support System (IDSS)

From the 70's [Little, 2004] the field of *Decision Support Systems* (DSS) is responsible for the generation of software that can analyze data and provide answers to pertinent and relevant questions in making decisions of a specific organization. Initially, the DSS were responsible to deal with some basic monitoring and analysis data. Later, the model-based simulations allowed the *what-if* analysis, and eventually the DSS were incorporating more high level skills. By the late 90s, Intelligent Decision Support Systems (IDSS) [Marakas, 2003] begin to include specific domain knowledge and the ability of automatic reasoning. Research to establish a stable infrastructure to the rapid development of IDSS [Power, 2008] is intense, but there are still many open problems. However, it seems clear that the IDSS have to combine data-driven models, analytical models and knowledge-based models, and some capacity of reasoning to provide a relevant end-user support [Poch et al., 2004].

One of the main components of a module IDSS is the use of data, which may include methods from the data mining (DM) field, but also classical machine learning or multivariate statistics [Han et al., 2011]. This is the module that explores the underlying database system for extract the information that will become knowledge for the supporting the posterior decisions. As already been said, this proposal aims to use advanced automatic classification techniques [Gibert et al., 2014] and statistics that capture the complex structure of the study data, for supporting further decision-making related to the success of the pre-post intervention studies.

2.11 Summary

This chapter contains the needed concepts to understand the present thesis and their background in the concerned areas. It is remarkable that techniques of AI or Data Mining are minimally applied to the data of clinical assays such as the intervention studies. Through the research in the literature, this lack of interdisciplinarity is found between both clinical intervention studies and AI. Most of the data from this type of studies is analyzed using classical statistic methods.

A characteristic in this type of studies is the big amount of attributes. Usually, the data is heterogeneous mixing numerical and categorical attributes. In addition, there are attributes with different roles when more information about the individuals such as the medical records apart from the gene expression are provided.

There is a need of new methodologies for analyzing the intervention effects when environmental factors that can interact with the intervention itself are not all well-known.

One of the problems is the difficulty of finding available data of this type of studies. The main reasons are that there are not many studies yet due to the high cost both in time and money and usually, these studies are not public. Most works in genetics do not analyze gene expressions with other physiological or environmental data.

A frequent comment is the need of techniques allowing an easy discovery of the multivariate interactions between factors. For that reason, we think that clustering can help in this aspect. Clustering can identify multivariate patterns in a more natural way without requiring previous knowledge. Finding a prior substructure of the problem can make easier to get richer models.

The validation of the resulting clusters is widely approached using the cluster validity indexes. These indexes evaluate the clusters from a structural point of view.The definition of the most used CVIs with common notation allows classifying them depending on what they evaluate: cohesion inside each cluster, separability between different clusters or a relation between these two criterion. Nevertheless, these indexes can not evaluate the meaning of these clusters. Thus, the structural validity does not ensure the usefulness of clustering, as meaningfulness is also a key feature to guarantee de decision-making support.

Therefore, cluster interpretation is an important part of the data mining process and can help in the transference of knowledge from the AI experts to clinical experts. This area is an open problem due to the difficulty of proving that the results are correct.

From the analysis of this literature, the need for new approaches to analyze this type of studies is contrasted. In next Chapter 3, a new methodology is proposed to handle the analysis of the intervention pre-post studies. This proposed methodology is based on clustering techniques to find similar subgroups of individuals and the use of a new cluster interpretation methodology to characterize the resulting groups.

Chapter 3

Thesis Proposal

This chapter is the core of this PhD Thesis. In this chapter, the problem to be solved and the methodology that we propose to approach this problem are defined.

In this thesis, we want to approach the analysis of pre-post studies using artificial intelligence techniques. Usually, this type of studies contains a large number of attributes that can have different origins. The individuals are assigned to one intervention group (commonly including a control group). These studies are characterized by having measures of the attributes before and after the intervention. In Section 3.1, this model is enriched differentiating between those attributes that does not change during the intervention and those that change. Also, we differentiate between two types of attributes experimenting changes: those with changes that are more difficult to be detected and are approached with a multiplicative model and the other that are approached with an additive model.

Section 3.2 contains the proposed global methodology and in the subsequent sections, the elements composing this methodology are detailed. The main elements that composed the proposed methodology are the use of an Integrated Multiview Clustering (detailed in Section 3.2.1), a generic cluster interpretation methodology (detailed in Section 3.6), a cluster interpretation methodology that copes with nested partitions (detailed in Section 3.7) and the characterization of trajectories between two partitions depending on the assigned intervention (see Section 3.10).

3.1 Formal Description of the Problem

Given a set of individuals $\mathcal{I} = \{i_1, ..., i_n\}$ and:

- $\mathcal{X} = \{X_1, ..., X_K\}$ the set of intrinsic characteristics of individuals that do not change during the study.

- Matrix $\chi = \begin{pmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nK} \end{pmatrix}$ where cells x_{ik} contain the value of attribute X_k for the individual $i \in \mathcal{I}$.

- $\mathcal{Y} = \{Y_1, ..., Y_L\}$ the set of attributes that may change along the study describing the state of individuals.

- $\mathcal{Y}_o = \begin{pmatrix} y_{o_{11}} & \cdots & y_{o_{1L}} \\ \vdots & \ddots & \vdots \\ y_{o_{n1}} & \cdots & y_{o_{nL}} \end{pmatrix}$ the matrix with observations of $\mathcal{Y}$ vector for the n individuals before the intervention.

- $\mathcal{Y}_f = \begin{pmatrix} y_{f_{11}} & \cdots & y_{f_{1L}} \\ \vdots & \ddots & \vdots \\ y_{f_{n1}} & \cdots & y_{f_{nL}} \end{pmatrix}$ the matrix with observations of $\mathcal{Y}$ vector for the n individuals after the intervention

- T a qualitative attribute indicating the kind of treatment or intervention assigned to the individual. Considering t different treatments, the values of T are $\mathcal{D}_T = \{T_1, ..., T_t, \mathcal{C}\}$ being $\mathcal{C}$ the control group.

- $\Delta = \left(\mathcal{Y}_f - \mathcal{Y}_o\right)$ $=(\Delta_1, ..., \Delta_L)$ the matrix containing the effect of the intervention on $\mathcal{Y}$ attributes, composed by L attributes where $\Delta_\ell = \mathcal{Y}_{f_\ell} - \mathcal{Y}_{o_\ell}$, $\ell = 1 : L$ and *measures the change of attribute $\mathcal{Y}_\ell$ during the intervention. Here an assumption of additive effect is assumed.*

- $\mathcal{Z} = \{Z_1, ..., Z_M\}$ the set of relevant covariables that can change during the study and *have also a small and local effect on $\mathcal{Y}$ attributes, not observable in global models.*

- $Z_o = \begin{pmatrix} z_{o_{11}} & \cdots & z_{o_{1M}} \\ \vdots & \ddots & \vdots \\ z_{o_{n1}} & \cdots & z_{o_{nM}} \end{pmatrix}$ the matrix with observations of $\mathcal{Z}$ vector for the n individuals before the intervention.

- $Z_f = \begin{pmatrix} z_{f_{11}} & \cdots & z_{f_{1M}} \\ \vdots & \ddots & \vdots \\ z_{f_{n1}} & \cdots & z_{f_{nM}} \end{pmatrix}$ the matrix with observations of $\mathcal{Z}$ vector for the n individuals after the intervention.

- $\Lambda = (\Lambda_1, ..., \Lambda_M)$ the matrix containing the *multiplicative effect of the intervention over $\mathcal{Z}$ attributes.* Thus $\Lambda_m = \frac{\mathcal{Z}_{mf}}{\mathcal{Z}_{mo}}$, $m = 1 : M$

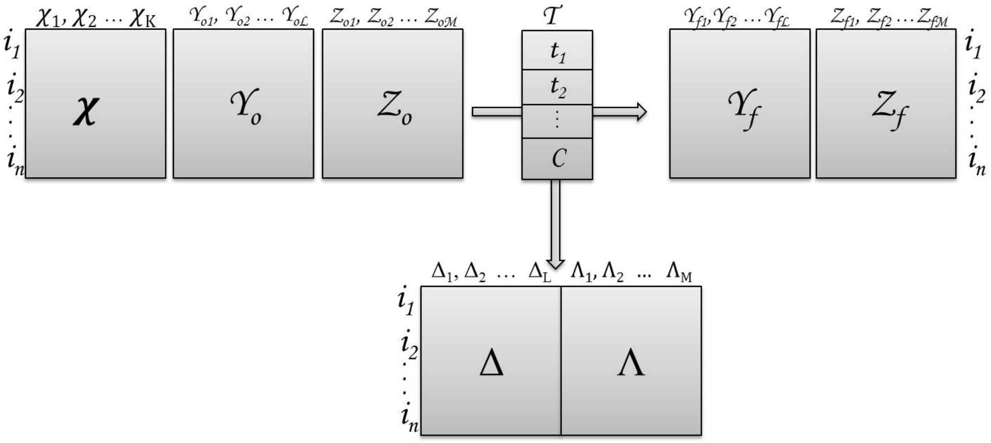

$\chi_1, \chi_2 \dots \chi_K$
$\Upsilon_{o1}, \Upsilon_{o2} \dots \Upsilon_{oL}$
$Z_{o1}, Z_{o2} \dots Z_{oM}$
T
$\Upsilon_{f1}, \Upsilon_{f2} \dots \Upsilon_{fL}$
$Z_{f1}, Z_{f2} \dots Z_{fM}$
i_1
i_2
i_n
χ
Υ_o
Z_o
t_1
t_2
C
Υ_f
Z_f
$\Delta_1, \Delta_2 \dots \Delta_L$
$\Lambda_1, \Lambda_2 \dots \Lambda_M$
Δ
Λ

Figure 3.1: Structure of the PhD.thesis-problem

The problem to be analyzed in this PhD. Thesis is the following:

Find a methodology for finding a comprehensible model to understand both:

- *the joint effect of T and $\mathcal{X}$ over $\mathcal{Y}$ (by using Δ).*
- *the adherence to the proposed intervention T to each individual*

Synthesize the joint effect of T and $\mathcal{X}$ over $\mathcal{Y}$ in a set of prototypical effects $E = \{e_1, \dots, e_{n_E}\}$ described in terms of distinguishable patterns of Δ.

Find how $\mathcal{Z}$ is associated with each type of effect $e \in E$ by analyzing local relationships between $\mathcal{Z}$ & ($\mathcal{X}|_e$, $\mathcal{Y}|_e$, $T|_e$), $e \in E$ (by using Λ).

3.2 Methodology for Analyzing Pre-Post Data

Nowadays, results obtained from classical pre-post studies are based on traditional and often basic, statistical techniques (see Chapter 2). Till now, these techniques have not been expressive enough to allow the extraction of complex relationships, as the level and degree of interactions between different subset of attributes of different types and distributions is too complex, in this context, to be captured by simple data analysis or pre-post statistical testing.

We propose a methodology for analyzing this type of studies which include a combination of AI and statistical techniques and decompose the problem in such a way that local multivariate interactions could be detected and modeled.

Clustering methods will be introduced to reduce the original problem space on rows dimensionality to a simpler set of more abstract elements to be modeled, each one concentrating more general conceptual contents. On the other hand, the attributes space will be also split in different subsets playing different roles in the analysis according to their different structure and behavior.

The proposed Integrative Multiview Clustering Methodology for Pre-Post Studies can be described as follows.
Given

- $\mathcal{I} = \{i_1, ..., i_n\}$ the set of individuals
- $\mathcal{A}$ a generic matrix containing all the attributes describing the individuals of $\mathcal{I}$
- $\mathcal{X} \subseteq \mathcal{A}$ the set of intrinsic attributes that do not change along the study,
- $\mathcal{Y} \subseteq \mathcal{A}$ the set of attributes that may changes along the study,
- χ the data matrix of $\mathcal{X}$,
- $\mathcal{Y}_o$ the data matrix of the characteristics describing the initial state of individuals before the intervention that can change during the study,
- $\mathcal{Y}_f$ the corresponding data matrix of measures of $\mathcal{Y}$ performed after the intervention,
- T the intervention assigned to the individuals,
- Δ the matrix containing the additive effect of the intervention on $\mathcal{Y}$ attributes,
- $\mathcal{Z}$ the set of relevant covariables that can change during the study and have also a small and local effect on $\mathcal{Y}$,
- $\mathcal{Z}_o$ measured before the intervention,
- $\mathcal{Z}_f$ are measured after the intervention and
- $\mathcal{D}$ the matrix containing the multiplicative effect over $\mathcal{Z}$ attributes.

1. Data Preprocessing: real data require a preprocessing step to prepare them for the analysis.

2. Create a small number of *Thematic Blocks*:

   (a) Divide $\mathcal{A} = \mathcal{X} \cup \mathcal{Y}$ attributes in a small number of thematic blocks using the background expertise in the area: $\mathcal{B} = \{\mathcal{B}^1, \mathcal{B}^2, \ldots, \mathcal{B}^\beta\}$ such that:

   - $\bigcup\limits_{\forall b \in \mathcal{B}} b \subseteq \mathcal{X} \cup \mathcal{Y}$
   - $\forall b, b' \in \mathcal{B} : b \cap b' = \emptyset$

   (b) Build $\mathcal{B}_o^1,\ \mathcal{B}_o^2, \ldots,\ \mathcal{B}_o^\beta$ and $\mathcal{B}_f^1,\ \mathcal{B}_f^2, \ldots,\ \mathcal{B}_f^\beta$ accordingly.

3. Profiling the Initial state of individuals:

   (a) Build an Integrative Multiview clustering over the blocks $\mathcal{B}_o^1,\ \mathcal{B}_o^2, \ldots,\ \mathcal{B}_o^\beta$ getting an integrated partition $\mathcal{P}_o$ (see Section 3.2.1).

   (b) Characterize, interpret and consequently label the classes of $\mathcal{P}_o$ using the interpretations of $\mathcal{P}_{o_1}, \mathcal{P}_{o_2}, \ldots, \mathcal{P}_{o_\beta}$ with the Cluster Interpretation Methodology NCI-IMS proposed in Section 3.9.

4. Profiling the Final state of individuals
   (a) Build an Integrative Multiview clustering over the blocks $\mathcal{B}_f^1$, $\mathcal{B}_f^2, \dots,$ $\mathcal{B}_f^\beta$ getting an integrated partition $\mathcal{P}_f$ (see Section 3.2.1):
   (b) Characterize, interpret and consequently label the classes of $\mathcal{P}_f$ using the interpretations of $\mathcal{P}_{f_1}, \mathcal{P}_{f_2}, \dots, \mathcal{P}_{f_\beta}$ with the Cluster Interpretation Methodology NCI-IMS proposed in Section 3.9.

5. Trajectory Analysis:
   - Build the Pre-Post Trajectories Map of $\mathcal{P}_o \times T$ and $\mathcal{P}_f$ (see Section 3.10)
   - Interpret the characteristic paths between initial states $\mathcal{P}_o \times T$ and final states $\mathcal{P}_f$ in terms of classes labels previously build.

6. Analysis of adherence to the Intervention (T) (when possible)
   (a) Identify $\mathcal{Y}^T$, the subset of attributes in $\mathcal{Y}$, directly related with adherence to intervention
   (b) Analysis of the adherence to the intervention T over $\mathcal{P}_o$:
      - Build the Table of profiles comparisons between $\mathcal{Y}_o^T|\mathcal{P}_o \times T$ versus $\mathcal{Y}_f^T|\mathcal{P}_o \times T$
      - Analyze whether differences in profiles reflect adherence to prescribed interventions or not.
   (c) Analysis of the adherence to the intervention T over $\mathcal{P}_o \times \mathcal{P}_f$:
      - Build the table of profiles comparisons between $\mathcal{Y}_o^T|\mathcal{P}_o \times \mathcal{P}_f \times T$ versus $\mathcal{Y}_f^T|\mathcal{P}_o \times \mathcal{P}_f \times T$
      - Synthesize the relationships among $\mathcal{P}_o, \mathcal{P}_f$ and T in terms of changes in profiles and adherence to the intervention and associate to trajectories in the Pre-Post Trajectory Map.

7. Analysis of the Intervention effects E using the characterization of the Pre-Post Trajectories Map obtained in Step 5.
   (a) Characterize the additive effect Δ of the intervention T depending on each trajectory: Build the table of profiles comparisons between $\mathcal{Y}_o|\mathcal{P}_o \times \mathcal{P}_f \times T$ versus $\mathcal{Y}_f|\mathcal{P}_o \times \mathcal{P}_f \times T$.
   (b) Characterize the multiplicative effect Λ of the intervention T depending on each trajectory: Build the table of profiles comparisons between $log(\mathcal{Z}_o)|\mathcal{P}_o \times \mathcal{P}_f \times T$ versus $log(\mathcal{Z}_f)|\mathcal{P}_o \times \mathcal{P}_f \times T$.
   (c) Build Models $\Lambda|_e = f\big(\chi|_e, \mathcal{Y}_o|_e, \Delta|_e, T|_e, \mathcal{Z}|_e\big)$

3.2.1 Integrative Multiview Clustering

Integrative Multiview Clustering (IMC) is proposed to find patterns on datasets with a big number of attributes suitable to be divided in thematic blocks (see Figure 3.2). The idea is to split the original set of attributes into different subsets that can be analyzed locally according to the semantics of the attributes, as in original Multiview Clustering approach. Here the obtained partitions are proposed to be integrated together in a single partition crossing all local ones.

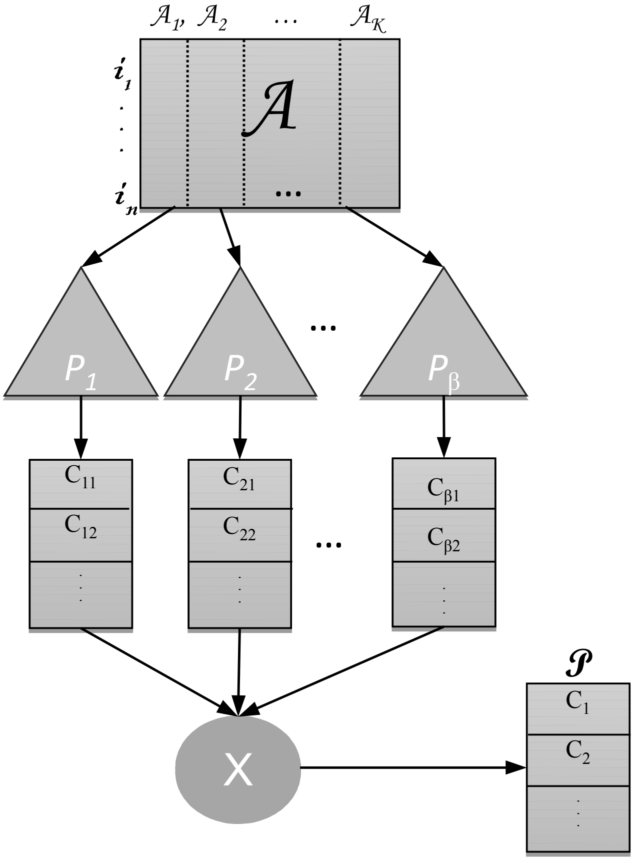


Figure 3.2: **Integrative Multiview Clustering Scheme**. Data Matrix A is split into β different thematic blocks of attributes. For each block, a clustering is performed and the β partitions are crossed giving as a result the final partition P.

Given $\mathcal{I} = \{i_1, ..., i_n\}$ the set of individuals and $\mathcal{B}^1$, $\mathcal{B}^2, \ldots,$ $\mathcal{B}^\beta$ the thematic blocks, the Integrative Multiview Clustering is built as following:

1. Clustering of each block $\mathcal{B}^1, \ldots, \mathcal{B}^\beta$. This produces several partitions of $\mathcal{I} : \mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_\beta$

2. Build an integrated partition: Build the cross clustering

   $\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2 \times \cdots \times \mathcal{P}_\beta$

Any clustering method can be used provided that it is the same in all the Blocks.

The domain knowledge is used for identifying the different roles of the attributes and used to define the thematic blocks. These blocks are made in such a way that each one regards a different aspect of the individuals. The complexity of combining partial analysis on every block is significantly lower than the global analysis.

However, giving several disconnected views of the data do not help too much to decision-making in the particular context of study. Thus, our proposal is to extend the Multiview Clustering approach by an *integrative step* where the results of each view are properly combined to give a single partition taking into account the structure found under all views. Our proposal consists in using the Cartesian Product of all partitions obtained under each view.

An important property of this proposal is that the resulting clustering P is embedded or nested in partitions used to build it $P_1, ...$ and P_β (see more details about nested partitions in Section 3.7).

The resulting classes of the IMC are more interpretable as it can be seen in Section 6.2.

In this thesis, the IMC approach is built using hierarchical clustering because the number of clusters is not known and using hierarchical methods can be induced as an output of the method. Ward 's minimum variance method has been used as it use ti provide more interpretable classes than others (see Section 2.3.1). This clustering method requires a distance in the inner kernel between objects and since the data is heterogeneous a distance that handles messy data is needed. The Gower's distance is selected for this reason (see Section 2.3). In Section 6.1, this option is compared against other clustering methods from the structural point of view concluding that Ward's method is suitable for finding compact clusters.

3.3 Assessing the Significance of an Attribute versus a Partition

Once data is partitioned into classes by P, the understanding of classes requires analyzing which attributes show significant differences among classes.

Classic Approach: Attribute X shows significant differences respect to the set of classes P:

- if X is numerical: use ANOVA or Kruskal-Wallis test (see definition in Section 2.9.1), depending on properties of X itself.

- if X is qualitative: use χ^2-Independence test (see definition in Section 2.9.2)

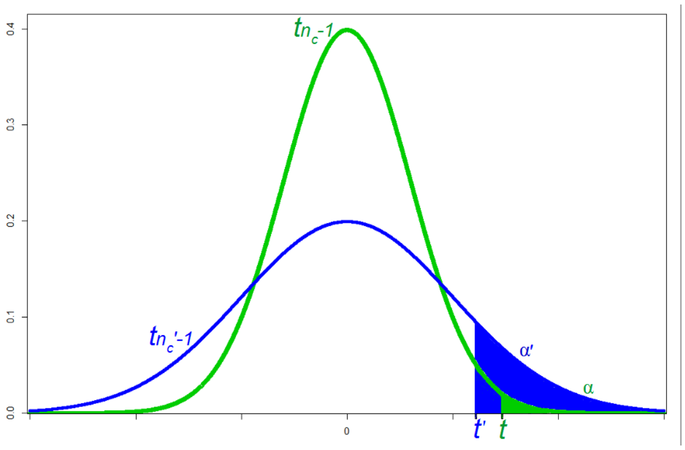


Figure 3.3: Overview of Property 1

If the corresponding test is significant, the attribute X is behaving different at least in one class.

However, these tests do not provide information about which class or classes is/are different nor about the sense of the differences. Multiple comparisons tests should be used afterwards to identify these differences. However, the process is combinatorial and an alternative procedure is proposed in this thesis.

From the Multivariate Analysis field [Lebart et al., 2000], we import more specific tests to assess the significance of an attribute to every single class. *Test-Value* (as described in Section 2.9.3) is used to this purpose in Multivariate Analysis. In Section 6.3, it is shown that the Test-Value subsume other classical statistical tests making the Test-Value a suitable test for Cluster Interpretation.

However, in our context, using the Test-Values in their original form might produce inconsistencies at the interpretation level.

Since same local means might appear significant or not in different classes depending on the size of the class.

This occurs because the *Test-Value* uses the variance of the mean in the denominator to determine the value of the statistics, and those variances are always inversely proportional to the sample sizes, in such a way that the same mean differences between class mean and global sample mean lose significance when sample size is reduced.
Given

- $\mathcal{I} = \{i_1, ..., i_n\}$ a set of individuals,
- X an attribute with n observations $\{x_1, \ldots, x_n\}$. Attribute X might be either numerical or qualitative, let $\mathcal{D}_X$ be the set of possible values for X.
- $P = \{C_1, \ldots, C_\xi\}$ a partition of $\mathcal{I}$ and $\forall C \in P : n_c = card(C)$

Property 1: $\bar{X}^c - \bar{X}$ loses significance when n decreases under Test-Value:

Let $E(X) = \mu, V(X) = \sigma^2$ then $V(\overline{X}) = \frac{\sigma^2}{n}$.

Let $\hat{\mu} = \overline{X}, \hat{\sigma}^2 = s^2$ then $V(\overline{X})$ is estimated by $\frac{s^2}{n}$.

Let $X^c = X|P = C$ the X attribute conditioned to class C with $E(X^c) = \mu_c$ and $\hat{\mu}_c = \overline{X}^c$

Since X^c emerges from C, and $C \subset I$, and considering that the null hypothesis of all the significance tests that we are assessing assumes equality, under H_0 it can be assumed that $V(X) = V(X^c) = \sigma^2$

And under this premise, it holds that $V(\overline{X}_c) = \frac{\sigma^2}{n_c}$, estimated by $\frac{s^2}{n_c}$.

Let us consider two samples I and I' of sizes $n' < n$ and a attribute X' measured over I', where $E(X') = E(X)$ and $V(X') = V(X)$; let us also consider two classes $C \subset I$ and $C' \subset I'$ such that $\pi_{c'} = \pi_c$, then $\frac{n_{c'}}{n'} = \frac{n_c}{n}$ and this implies $n'_c < n_c$

$V(\overline{X}^{c'})$ is estimated by $\frac{s^2}{n_{c'}}$. Thus $V(\bar{X}^{c'}) > V(\bar{X}^c)$

It also holds that $\left(1 - \frac{n_{c'}}{n'}\right) = (1 - \pi_{c'}) = (1 - \pi_c) = \left(1 - \frac{n_c}{n}\right)$

And this means that the denominator of the statistics:

$\left(1 - \frac{n_{c'}}{n'}\right) \frac{s^2}{n_{c'}} = \left(1 - \frac{n_{c'}}{n'}\right) V(\bar{X}^{c'}) > \left(1 - \frac{n_c}{n}\right) V(\bar{X}^c) = \left(1 - \frac{n_c}{n}\right) \frac{s^2}{n_c}$

Moreover, $\bar{X}^{c'} - \bar{X}' = \bar{X}^c - \bar{X}$ by construction

And then, given $t' = \frac{\bar{X}^{c'} - \bar{X}'}{\sqrt{(1 - \frac{n'_c}{n'}) \frac{s^2}{n'_c}}} \sim t_{n'_c - 1}$ and $t = \frac{\bar{X}^c - \bar{X}}{\sqrt{(1 - \frac{n_c}{n}) \frac{s^2}{n_c}}} \sim t_{n_c - 1}$ and having both statistics equal numerator, it holds that $t' < t$

As the p-value of the test computes the tail of the Student's T distribution,

$$\alpha' = P(t_{n_{c'} - 1} \geq t' > P(t_{n_c - 1} \geq t) = \alpha)$$

and the significance of $\bar{X}^c - \bar{X}$ results higher than the significance of the same quantity for a smaller sample (see Figure 3.3). (qed)

This produces that the same difference between global and local mean is significant (and should appear in the description of the class) in a bigger class and non significant in smaller one (where the attribute should not appear as a descriptor of the class)

It is obvious that for interpretation purposes this situation is not acceptable. For this reason, a modification of the classical *Test-Value* is proposed in this work, based on the idea that all classes come from the same referent population under H_o and for all of them global variance will be used under H_o. See [Gibert et al., 2015] for details about the needs of this contribution and in Section 6.4 for the experimental details.

3.3.1 Modified *Test-Value*

Thus, the new *Modified Test-Value* is:

Let X a numerical attribute and $X^c = X|C$,

$$\begin{array}{l} H_o : \mu_c = \mu \\ H_1 : \mu_c \neq \mu \end{array} \qquad \tau = \frac{\bar{X}^c - \bar{X}}{\sqrt{(1 - \frac{n_c}{n})\frac{s^2}{n}}} \sim t_{n-1} \tag{3.1}$$

Let X a qualitative attribute, $s \in \mathcal{D}_X$, $\mathcal{D}_X$ the set of possible values for X,

$$\begin{array}{l} H_o : \pi_c = \pi \\ H_1 : \pi_c \neq \pi \end{array} \qquad \pi = \frac{\frac{n_{sc}}{n_c} - \frac{n_s}{n}}{\sqrt{(1 - \frac{n_c}{n})\frac{\frac{n_s}{n}(1 - \frac{n_s}{n})}{n}}} \sim z \tag{3.2}$$

This may be also be written as,

$$\pi = \frac{p_{sc} - p_s}{\sqrt{(1 - \frac{n_c}{n})\frac{p_s(1-p_s)}{n}}} \sim z$$

where,

$p_s = \frac{card(i \in X : x_i = s)}{n} = \frac{n_s}{n}$ the empirical probability of the value s in the global sample
$p_{sc} = \frac{card(i \in C : x_i = s)}{n_c} = \frac{n_{sc}}{n_c}$ the empirical probability of the value s in class C.

3.3.2 Assessment in Pre-Post Scenarios

Classical pre-post studies use t-test for paired samples to assess the effects between instant t and $t-1$. Eventually this is done by taking into account the several treatments occurred between t and $t-1$. This is equivalent to partition individuals according to treatment to be followed and to perform the paired t-test locally to each treatment group.

This approach assumes that the partitions in both t and $t-1$ instants are the same, the one induced by the different treatments; Also, homogeneity among samples is assumed inside every treatment.

Nevertheless, in the present scenario, homogeneity is not assumed and the referent partition used is not the prescribed diet, but the dietary profile before the intervention and the dietary profile after it. Thus, the partition in t in general will not be the same as the one in instant $t-1$. Thus making difficult the use of the classical t-test approach.

For these reasons, the tests τ^t and π^t are introduced in order to assess the difference of an attribute along time when partitions are not the same in the two instants compared.

To assess the significances of an attribute in a class at instant t w.r.t. a previous situation (in pre-post analysis scenarios, for example, or in temporal studies, in general), we introduce a new way to make the test, by taking as a reference not the total sample mean at instant t, but that of $t-1$, which assesses the effect over X of passing from $t-1$ to t:

Let X a numerical attribute, X_{t-1} and X_t are its corresponding measures in time $t-1$ and t,

$$\begin{array}{l} H_o : \mu_{c_t} = \mu_{c_{t-1}} \\ H_1 : \mu_{c_t} \neq \mu_{c_{t-1}} \end{array} \qquad \tau^t = \frac{\bar{X}_t^c - \bar{X}_{t-1}}{\sqrt{(1-\frac{n_c}{n})\frac{s_{t-1}^2}{n}}} \sim t_{n-1} \tag{3.3}$$

Let X a qualitative attribute, X_{t-1} and X_t are its corresponding measures in time $t-1$ and t and s the category to treat,

$$\begin{array}{l} H_o : \pi_{c_t} = \pi_{c_{t-1}} \\ H_1 : \pi_{c_t} \neq \pi_{c_{t-1}} \end{array} \qquad \pi^t = \frac{\frac{n_{sc_t}}{n_c} - \frac{n_{s_{t-1}}}{n}}{\sqrt{(1-\frac{n_c}{n})\frac{\frac{n_{s_{t-1}}}{n}(1-\frac{n_{s_{t-1}}}{n})}{n}}} \sim z \tag{3.4}$$

Which is equivalent to,

$$\pi^t = \frac{p_{sc_t} - p_{s_{t-1}}}{\sqrt{(1-\frac{n_c}{n})\frac{p_{s_{t-1}}(1-p_{s_{t-1}})}{n}}} \sim z$$

From now on Θ will be used as a generic statistics for any of these relevant tests, $\Theta = \{\tau, \pi, \tau^t, \pi^t\}$.

Both tests τ^t and π^t measure the joint effect of the previous dietary profile together with the intervention followed. It can be shown that the proposed τ^t and π^t tests are equivalent to the sum of both the t-test assessing the effect of the dietary profile w.r.t global sample and the paired t-test assessing the effect of intervention local to every dietary profile.

In a classical pre-post analysis the t-test

$$\begin{array}{l} H_o : \mu_c = \mu_o \\ H_1 : \mu_c \neq \mu_o \end{array} \qquad \frac{\bar{X}^c - \mu_o}{\sqrt{\frac{s_c^2}{n_c}}} \tag{3.5}$$

assesses the homogeneity between the classes induced by an intervention in the pre step and global sample.

On the other hand,

$$\begin{array}{l} H_o : \mu_{c_{t-1}} = \mu_{c_t} \\ H_1 : \mu_{c_{t-1}} \neq \mu_{c_t} \end{array} \qquad \frac{(\bar{X}_t^c - \bar{X}_{t-1}^c) - (\mu_t - \mu_{t-1})}{\sqrt{\frac{s_{c_{t-1}}^2}{n_c}}} \tag{3.6}$$

assesses the effect of the intervention in group C. Our proposal is based on the idea that

$H_o : \mu_{c_t} = \mu_o$
$H_1 : \mu_{c_t} \neq \mu_o$

$$\frac{\bar{X}_{t-1}^c - \mu_{t-1}}{\sqrt{\frac{s_{c_{t-1}}^2}{n_c}}} + \frac{(\bar{X}_t^c - \bar{X}_{t-1}^c) - (\mu_t - \mu_{t-1})}{\sqrt{\frac{s_{c\Delta}^2}{n_c}}} = \frac{\bar{X}_t^c - \mu_{t-1}}{\sqrt{\frac{s_{c_{t-1}}^2}{n_c}}}$$

Assuming that under H_o the attribute does not change $(V(X)_t = V(X)_{t-1})$ and the addition of both statistics provides the proposed test:

This proposal assesses whether group C after the intervention shows significant difference w.r.t. global initial levels.

Following the same approach a similar result is obtained by the Test-Value in both the case of quantitative and qualitative attributes.

It can be shown that the proposed τ^t and π^t tests are equivalent to the sum of both *Test-Value* assessing the effect of the dietary profile and *Test-Value* assessing local effect of intervention.

The τ test

$H_o : \mu_c = \mu_o$
$H_o : \mu_c \neq \mu_o$

$$\frac{\bar{X}^c_{t-1} - \bar{X}_{t-1}}{\sqrt{(1-\frac{n_c}{n})\frac{s^2_{t-1}}{n}}}$$

assesses the homogeneity between classes induced by several interventions in the pre step.

On the other hand

$H_o : \mu_{c_{t-1}} = \mu_t$
$H_1 : \mu_{c_{t-1}} \neq \mu_t$

$$\frac{(\bar{X}^c_t - \bar{X}^c_{t-1}) - (\bar{X}_t - \bar{X}_{t-1})}{\sqrt{(1-\frac{n_c}{n})\frac{s^2_{t-1}}{n}}}$$

assesses the effect of the intervention in group C. Assuming that under H_o the attribute does not change ($V(X)_t = V(X)_{t-1}$), the addition of both statistics provides the proposed test:

$H_o : \mu_{c_t} = \mu_o$
$H_1 : \mu_{c_t} \neq \mu_o$

$$\frac{\bar{X}^c_{t-1} - \bar{X}_{t-1}}{\sqrt{(1-\frac{n_c}{n})\frac{s^2_{t-1}}{n}}} + \frac{(\bar{X}^c_t - \bar{X}^c_{t-1}) - (\bar{X}_t - \bar{X}_{t-1})}{\sqrt{(1-\frac{n_c}{n})\frac{s^2_{\Delta}}{n}}} = \frac{\bar{X}^c_t - \bar{X}_{t-1}}{\sqrt{(1-\frac{n_c}{n})\frac{s^2_{t-1}}{n}}}$$

assesses whether group C after the intervention shows significant difference w.r.t. global initial levels. The qualitative approach π^t behaves equivalently.

3.4 Interpreting an Attribute by means of a Modified *Test-Value*

Cluster Interpretation is a post-process of finding the common and distinctive characteristics of every class and creating the corresponding profiles. However, cluster interpretation is still an open issue from the methodological point of view (see Section 2.6).

The proposed methodology for the interpretation of an attribute X in the classes of partition P is the following:

1. Determine δ: determine the threshold $\delta \in [0, 1]$ representing the admitted noise in the class.

2. Build $X|P = \{X|C_1, \ldots, X|C_\xi\}$, Name $X^c = X|P = C,\ C \in P$

3. Determine $\alpha \in [0,1]$ as a significance level

4. $\forall C \in P$,

   The descriptor of the class C according to the attribute X is defined as:

   $d_{X,C} = \{d_{W,C}\}$,

   where,

   $d_{W,C} = (W, C, sense)$, being:

   $$W = \begin{cases} "X" & X \textit{ is numeric and } X \textit{ is descriptor of } C \\ < s, X > & s \in \mathcal{D}_X,\ X \textit{ is qualitative and} \\ & s \textit{ is descriptor (or basic descriptor) of } C \end{cases}$$

   $sense = \{\top, \bot, \uparrow, \downarrow\}$

   To determine a (basic) descriptor of X do:

   - if X qualitative:

     (a) Detect *Basic descriptor* (B): To detect quasi-constant attribute in a class.

     - $\exists s \in \mathcal{D}_X :\ p_{sc} \geq (1-\delta)$
       The value s will be a basic descriptor of C according to attribute X:
       Build $d_{W,C} = (< s, X >, C, \top)$
     - $\neg\exists s \in \mathcal{D}_X :\ p_{sc} \geq (1-\delta)$, $\forall s \in \mathcal{D}_X$:
       $p_{sc} \leq \delta$ The value s will be a basic descriptor of C according to attribute X:
       Build $d_{W,C} = (< s, X >, C, \bot)$

     (b) $\forall s \in \mathcal{D}_X$ such that s is *non Basic descriptor*:

     - Detect *significant descriptors* using the *Modified Test-Value*:
       - ∗ using π statistics (Equation 3.2) in general case.
       - ∗ using statistic π^t (Equation 3.4) if a temporal frame is considered.

       i. if p-value $\leq \alpha \implies s$ is significant in class C at level α and s is descriptor of C.

       ii. For significant tests, use the sign of the statistic value to determine the sense of the difference, given $\Theta = \{\pi, \pi^t\}$ and Θ_o the observed value of Θ:

       - ∗ if $\Theta_o > 0 \implies X$ has significant higher proportion of s in C than in the global sample: Build $d_{W,C} = (< s, X >, C, \uparrow)$

      - $*$ if $\Theta_o < 0 \implies X$ has significant lower proportion of s in C than in the global sample: Build $d_{W,C} = (< s, X >, C, \downarrow)$

- if X numeric:
  - compute the *Modified Test-Value* of X^c:
    - $*$ using statistic τ (Equation 3.1) in general case
    - $*$ using statistic τ^t (Equation 3.3) if a temporal frame is considered.
  - if p-value $\leq \alpha \implies X$ is significant in class C at level α and X is descriptor of C.
  - For significant tests, use the sign of the statistics to determine the sense of difference, given $\Theta = \{\tau, \tau^t\}$ and Θ_o the observed value of Θ:
    - $*$ if $\Theta_o > 0 \implies X$ has significant higher values in class C than in the global sample: Build $d_{W,C} = (X, C, \uparrow)$
    - $*$ if $\Theta_o < 0 \implies X$ has significant lower values in class C than in the global sample: Build $d_{W,C} = (X, C, \downarrow)$
- Build the descriptor $d_{X,P} = \bigcup_{\forall C \in P} d_{X,C}$.

  Let $\mathcal{D}_W$ be the set of all W appearing in $d_{X,P}$.

3.5 Sensitive Analysis of *Modified Test-Value* (SA-MTV)

A new element is introduced in the interpretation procedure which enables the descriptors of a class to be characterized according to the stability of their significance to small or large variations of the sample size. The idea behind this analysis is that large differences between class means and global means will provide significant Modified Test-Value, even with drastic reductions in sample size, also embracing the small sizes of the classes. In this scenario, a consistent interpretation will be obtained; whereas smaller differences might present unstable results that provide significance, or not, depending on the sample size.

Different strategies have been designed and tested (see Section 6.4 and Section 6.6). The conclusions indicate that the Modified Test-Value with the Sensitive Analysis is the best option.

In order to assess the robustness of the Modified Test-Value, an analysis with different sample sizes is proposed. Accordingly:

Given an attribute X and a partition P:

1. Determine $\varepsilon_1,\ \varepsilon_2 \in (0, 0.5] \wedge \varepsilon_1 + \varepsilon_2 < 1$

2. Build a set of *control sample sizes* to asses stability of results

   $\mathcal{S} = \{n(1-\varepsilon_1-\varepsilon_2), n(1-\varepsilon_1), n, n(1+\varepsilon_1), n(1+\varepsilon_1+\varepsilon_2)\}$, For simplicity, rename as $\mathcal{S} = \{\nu_i | i \in [1, 5]\}$ (eventually $\mathcal{S}$ might contain a more fine griding of sample size space)

3. For a given attribute X and a partition P, assess the significance using the algorithm described in Section 3.4 with the corresponding generalization of the Θ tests described in Section 3.3 (see the generalized version in Table 3.1) by using all $\nu \in \mathcal{S}$ as parameters and the corresponding test $\Theta = \{\tau_\nu, \tau_\nu^t, \pi_\nu, \pi_\nu^t\}$. Get $d_{X,P,\nu}, \forall \nu \in \mathcal{S}$.

   This analysis might be performed with any statistical test and provides information about how stable is a significance, and in consequence, how reliable is a descriptor to be considered in the profiling description, and to be the basis of further decisions.

   Therefore, the Modified *Test-Value* is generalized to be used with different sample sizes, the new *Generalized Test-Value* is:

X	**Static**	**Dynamic**
Numeric	$\tau_\nu = \dfrac{\bar{X}^c - \bar{X}}{\sqrt{(1 - \frac{n_c}{n})\frac{s^2}{\nu}}} \sim t_{\nu-1}$ (3.7)	$\tau_\nu^t = \dfrac{\bar{X}_t^c - \bar{X}_{t-1}}{\sqrt{(1 - \frac{n_c}{n})\frac{s_{t-1}^2}{\nu}}} \sim t_{\nu-1}$ (3.8)
Qualitative	$\pi_\nu = \dfrac{p_{sc} - p_s}{\sqrt{(1 - \frac{n_c}{n})\frac{p_s(1-p_s)}{\nu}}} \sim z$ (3.9)	$\pi_\nu^t = \dfrac{p_{sc_t} - p_{s_{t-1}}}{\sqrt{(1 - \frac{n_c}{n})\frac{p_{s_{t-1}}(1-p_{s_{t-1}})}{\nu}}} \sim z$ (3.10)

Table 3.1: *Generalized Test-Value*

   In a temporal framework where repeated measures of X are available τ_ν and π_ν assesses significant differences between the class among the global population whereas τ_ν^t and π_ν^t are used to assess significance between X_t^c and X_{t-1}.

4. The interpretation of P according to the attribute X is

   $\mathcal{V}_{X,P} = \{\mathcal{V}_{W,C}, \forall C \in P\}$

   being: $\mathcal{V}_{W,C} = (W, C, dp, sense)$ where

   - $sense \in \{\top, \perp, \uparrow, \downarrow, \times\}$
   - dp is the $description{-}power$ of W in class C , $dp \in \Pi = \{\overline{R}, \overline{M}, \overline{W}, W, M, R, B\}$ (see Table 3.2, where ✓ indicates test was significant and × indicates test was no-significant):

     B : X is *basic* descriptor.

R : X is highly *significant with robustness* in such a way that even reducing n till $n(1-\varepsilon_1-\varepsilon_2)$ of sample size, X keeps significant.

M : X is *moderate significant* in such a way that reducing n till $n(1-\varepsilon_1)$, X keeps significant but with a reduction $n(1-\varepsilon_1-\varepsilon_2)$, X loses the significance.

W : X is *weak significant* in such a way that is significant with the observed sample size n but even with a small reduction of sample size (like $n(1-\varepsilon_1)$), X loses the significance. This could be considered of spurious significance.

$\overline{W}$: X is *non-significant* with the original sample size n, but with a small increment of $\varepsilon_1\%$ of the sample size will provide significance.

$\overline{M}$: X is *non-significant* with the original sample size n and remains non-significant with an increment of $\varepsilon_1\%$ of the sample size, but X becomes significant with an increment of $(\varepsilon_1+\varepsilon_2)\%$ of sample size.

$\overline{R}$: X is *non-significant* and remains non-significant – even with an increment of $(\varepsilon_1+\varepsilon_2)\%$ of the sample size.

		Sample size				
		ν_1	ν_2	ν_3	ν_4	ν_5
Description Power	**Symbol**	$n(1-\varepsilon_1-\varepsilon_2)$	$n(1-\varepsilon_1)$	n	$n(1+\varepsilon_1)$	$n(1+\varepsilon_1+\varepsilon_2)$
Robust Non-descriptor	$\overline{R}$	×	×	×	×	×
Moderate Non-descriptor	$\overline{M}$	×	×	×	×	✓
Weak Non-descriptor	$\overline{W}$	×	×	×	✓	✓
Weak Descriptor	W	×	×	✓	✓	✓
Moderate Descriptor	M	×	✓	✓	✓	✓
Robust Descriptor	R	✓	✓	✓	✓	✓
Basic Descriptor	B	B	B	B	B	B
✓: $p-value<\alpha$		test $\in\Theta$				

Table 3.2: Sensitive Analysis

5. $\forall W \in \mathcal{D}_W$,

- Given a sense $\jmath \in \{\top,\bot\}$ then:
  - if $\forall\nu \in \mathcal{S} : (W,C,\jmath) \ \in d_{X,P,\nu}$ build $\mathcal{V}_{W,C}=(W,C,B,\jmath)$
- Given a sense $\jmath \in \{\uparrow,\downarrow\}$ then:
  - if $\forall\nu \in \mathcal{S} : (W,C,\jmath) \ \in d_{X,P,\nu}$ build $\mathcal{V}_{W,C}=(W,C,R,\jmath)$
  - if $\forall\nu \in \mathcal{S}\setminus\{\nu_1\} : (W,C,\jmath) \in d_{X,P,\nu} \wedge (W,C,\jmath) \notin d_{X,P,\nu_1}$ build $\mathcal{V}_{W,C} = (W,C,M,\jmath)$
  - if $\forall\nu \in \{\nu_3,\nu_4,\nu_5\} : (W,C,\jmath) \ \in d_{X,P,\nu} \wedge \forall\nu \in \{\nu_1,\nu_2\}\ (W,C,\jmath) \notin d_{X,P,\nu_1}$ build $\mathcal{V}_{W,C}=(W,C,W,\jmath)$
  - if $\forall\nu \in \{\nu_4,\nu_5\} : (W,C,\jmath) \in d_{X,P,\nu} \wedge \forall\nu \in \{\nu_1,\nu_2,\nu_3\}\ (W,C,\jmath) \notin d_{X,P,\nu_1}$ build $\mathcal{V}_{W,C}=(W,C,\overline{W},\jmath)$

- if $\forall \nu \in \mathcal{S} \setminus \{\nu_5\} : (W, C, \jmath) \notin d_{X,P,\nu} \wedge (W, C, \jmath) \in d_{X,P,\nu_5}$ build $\mathcal{V}_{W,C} = (W, C, \overline{M}, \jmath)$
- if $\forall \nu \in \mathcal{S} : (W, C, \jmath) \notin d_{X,P,\nu}$ build $\mathcal{V}_{W,C} = (W, C, \overline{R}, \times)$

3.6 Class Interpretation based on Integrated Marginal Significance (CI-IMS)

Combining all the elements introduced before, the CI-IMS methodology is proposed to identify *descriptors of a class.* The proposed interpretation methodology CI-IMS for a set of attributes $\mathcal{X}$ and a partition P is the following:
Given the set of attributes $\mathcal{X}$ and a partition P

1. Determine δ: determine the threshold $\delta \in [0, 1]$ representing the admitted noise in the class.

2. Determine $\alpha \in [0, 1]$ as a significance level.

3. Determine $\varepsilon_1,\ \varepsilon_2 \in (0, 0.5] \wedge \varepsilon_1 + \varepsilon_2 < 1$.

4. Determine a minimum level of robustness $r \in \Pi$.

5. Build $\mathcal{V}_P = \bigcup\limits_{X \in \mathcal{X}} \mathcal{V}_{X,P}$ by applying $SA - MTV(X, P, \varepsilon_1, \varepsilon_2, \alpha, \delta)$ defined in Section 3.5.

6. The interpretation of P at level $r \in \Pi$ is $\Upsilon_{P,r}$: as the subset of $\mathcal{V}_{W,C}$ such that all descriptors have at least the level of robustness r:

$$\Upsilon_{P,r} = \{\mathcal{V}_{W,C} : dp_{W,C} \geq r, \mathcal{V}_{W,C} \in \mathcal{V}_P\}$$

7. $\forall \mathcal{V}_{W,C} = (W, C, dp, sense) : sense \neq \times$, use a regular expression to translate to a verbose sentence:

$$\begin{cases} [X]\ is\ [ts]\ in\ the\ group\ [C] & if\ W = X \\ The\ group\ [C]\ has\ [ts]\ proportion\ of\ [s]([X]) & if\ W =< s, X > \end{cases}$$

$X \in \mathcal{L}_{\mathcal{X}} = \{X, \forall X \in \mathcal{X}\}$

$ts \in \mathcal{L}_{sense} = \{higher, lower\}$,

$$ts = \begin{cases} higher & if\ sense \in \{\top, \uparrow\} \\ lower & if\ sense \in \{\bot, \downarrow\} \end{cases}$$

$C \in \mathcal{L}_P = \{C, \forall C \in P\}$

Examples:

$\mathcal{V}_{W,C}$	Generated Sentence
$(< men, gender >, "M", B, \top)$	The group M has a higher proportion of men (gender)
$(< yes, cancer >, "YM", B, \perp)$	The group YM has a lower proportion of yes (cancer)
$(age, "WM", R, \uparrow)$	Age is higher in the group WM
$(height, "YW", R, \downarrow)$	Height is lower in the group YW
$(systolic\ pressure, "WM", M, \uparrow)$	Systolic pressure is higher in the group WM
$(BMI, "YW", M, \downarrow)$	BMI is lower in the group YW
$(glucose, "M", W, \uparrow)$	Glucose is higher in the group M
$(weight, "WM", W, \downarrow)$	Weight is lower in the group WM
$(sP - selectin, "M", \overline{W}, \uparrow)$	sP-selectin is higher in the group M
$(age, "M", \overline{W}, \downarrow)$	Age is lower in the group M
$(<\geq 2week, bakery >, "YW", \overline{M}, \uparrow)$	The group YW has a higher proportion of ≥2week (bakery)
$(MOHTyrosol, "WM", \overline{M}, \downarrow)$	MOHTyrosol is lower in the group WM

8. Visualize the Class Panel Graphs using only the subset of attributes that appear in $\Upsilon_{P,r}$.

3.7 Nested Partitions

Given a set of individuals $\mathcal{I} = \{i_1, ..., i_n\}$, $\mathcal{I}$ can be clustered into any of the elements of $\mathcal{P}(\mathcal{I})$. Given H an indexed hierarchy of partitions: $H = \{P_1, P_2, P_\ell ... P_n\}$, where

$P_1 = \{\mathcal{I}\}, P_1 \in \mathcal{P}(\mathcal{I})$,

$P_n = \{\{i_1\}, \{i_2\}, ..., \{i_n\}\}, P_n \in \mathcal{P}(\mathcal{I})$,

$\forall \ell, \ell' \in [1, n] : \ell' > \ell$ it holds that $P_{\ell'}$ is nested (embedded) in P_ℓ (see Figure 3.4).

Given two partitions P_ℓ, $P_{\ell'}$ of $\mathcal{I}$, $P_{\ell'}$ is nested in P_ℓ if:

- $\ell' > \ell$
- $\forall \ell \in [1, n-1], \forall C \in P_\ell$

  $\exists S_C \subseteq P_{\ell'} : \bigcup C' \in S_C = C \wedge$

  $\forall C' \in P_{\ell'} \wedge C' \notin S_C : C' \cap C = 0$

Therefore, partition $P_{\ell'}$ is a refinement or specialization of P_ℓ because each class of P_ℓ is divided into different classes in $P_{\ell'}$. The information that can provide P_ℓ can be inferred from $P_{\ell'}$ and not vice-versa. There are many situations where the nested partitions can arise, for instance the decisions trees, hierarchical clustering, any two criteria to split data, etc.

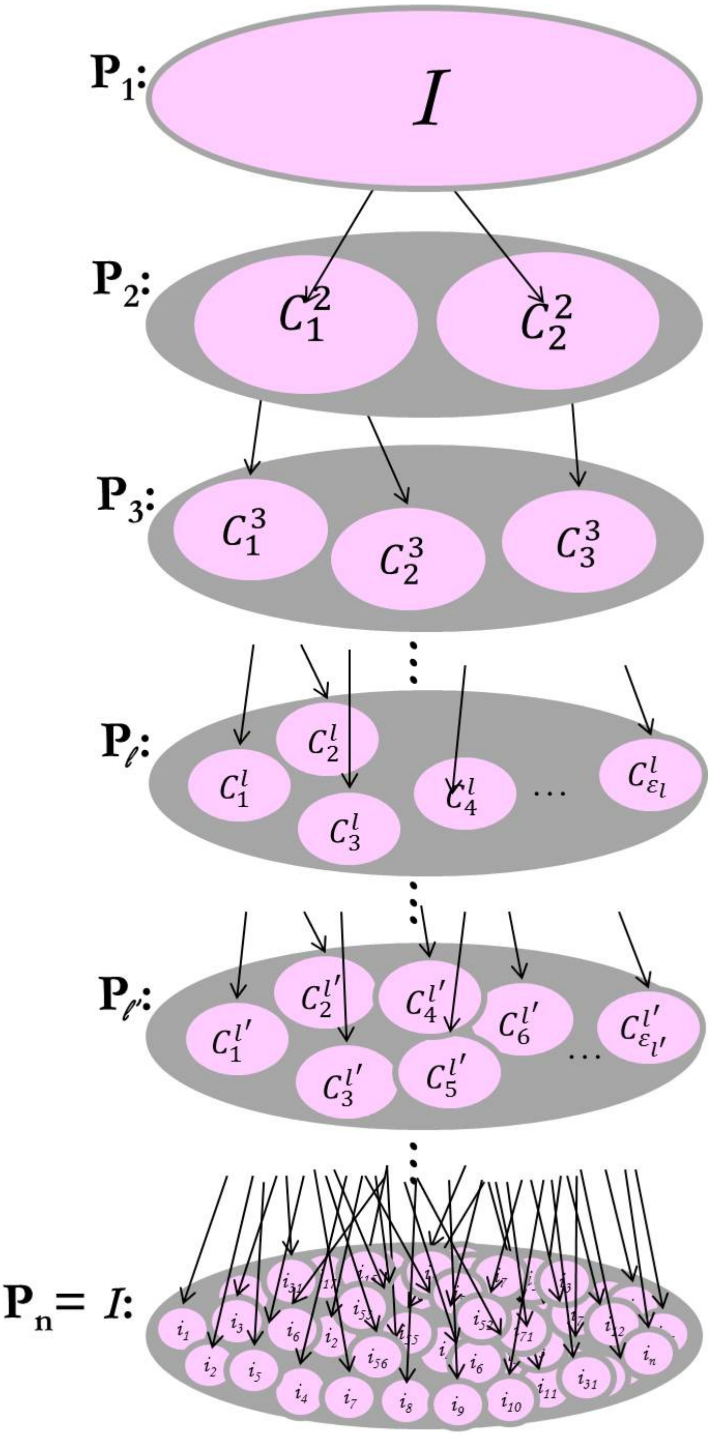


Figure 3.4: Hierarchy of Nested Partitions

3.8 Analyzing Consistency of a Descriptor between Nested Partitions (CNP)

The following situations could occur comparing the interpretations of two nested partitions: Given P_ℓ, $P_{\ell'}$ and $P_{\ell'}$ is nested in P_ℓ. Defining $S_C \subset P_{\ell'}$ as the set of subclasses of a class $C \in P_\ell$ (see Section 3.7). Then, given a descriptor W (which might refer either to a numerical attribute or to a category of a qualitative attribute), a class C and its set of subclasses S_C,

1. W is not significant in C and is significant in some subclasses ($W \neg significant\ in\ C \wedge \exists C' \in S_C : W significant\ in\ C'$). This represents *specialization*. The subdivision of C produces additional characteristics in some of the subclasses.

2. W is not significant in both C and all subclasses ($W \neg significant\ in\ C \wedge \forall C' \in S_C : W \neg significant\ in\ C'$). This represents the situation where the subdivision of C is not providing additional information and means that W is irrelevant.

3. W is significant in C and some subclasses ($W\ significant\ in\ C \wedge \exists C' \in S_C : W significant\ in\ C'$). This means that the characteristics of X have been *inherited* and propagated to some of the subclasses. Note that if W is significant for all subclasses, the subdivision of C does not provide additional information.

4. W is significant in C, but not significant in any subclass ($W\ significant\ in\ C \wedge \forall C' \in S_C : W \neg significant\ in\ C'$). This represents an *inconsistency*. It is not reasonable that the property that distinguishes the whole C from the global population does not appear significant for any of the subclasses of C.

Table 3.3: Diagnostic scenarios according to C and S_C interpretations

	SuperClass	
SubClasses	**Non-Significant**	**Significant**
Non-Significant	Irrelevant	Inconsistency
Significant	Specification	Inheritance

The 4 possible situations are synthesized in Table 3.3. With the diagnosis provided in Table 3.3 is possible to decide if it makes sense to use the descriptor W in the description of the superclass and that of the subclasses. Table $W - C$ (see Table 3.4) indicates, for the 4 diagnosis given in Table 3.3, whether the descriptor W must appear in the description of the superclass C and it can be seen that what is basically happening is that W describes C if it is significant in C and W do not describe C if it is not significant in C.

Table $W - C'$ (see Table 3.5) indicates how to proceed with the descriptor of $C' \in S_C$ in each diagnostic scenario. In fact, for a specific descriptor W and $C' \in S_C$, Table 3.5 reflects whether W becomes a descriptor of $C' \in S_C$ or not. In the case of W significant in C', it

Table 3.4: Inclusion of W as Class C descriptor according to diagnostic scenarios.

$(W - C)$	***Descriptor SuperClass***	
SubClass	**Non-Significant**	**Significant**
Non-Significant	No	Yes
Significant	No	Yes

Table 3.5: Inclusion of W as subclass C' descriptor according to diagnostic scenarios.

$(W - C')$	***SuperClass***	
Descriptor SubClass	**Non-Significant**	**Significant**
Non-Significant	None	None
Significant	Yes	Yes

appears in C' description. Table $W - C - C'$ (see Table 3.6) shows the situations resulting for the 4 diagnostic scenarios with respect to use W in the description of superclass C and one subclass $C' \in S_C$. Four different situation appear:

- None: irrelevant situation, W will not appear neither in C description nor in subclass $C' \in S_C$.
- SupC: inconsistent situation, W will appear in C but not in the description of the subclass $C' \in S_C$.
- Both: inheritance situation, W will appear both in C description and in the description of subclass $C' \in S_C$.
- SubC: specification situation, W will not appear in C description and it will appear in the description of the subclass $C' \in S_C$.

Table 3.6: Inclusion of W as superclass and subclasses' descriptor according to diagnostic scenarios.

$(W - C - C')$	***Descriptor SuperClass***	
Descriptor SubClass	**Non-Significant**	**Significant**
Non-Significant	None	SupC
Significant	SubC	Both

Table $W - S_C$ (see Table 3.7) analyses what happens with all subclasses in S_C together. In this case it is seen that when all subclasses behave homogeneously non significant, W does not appear in the description of none of S_C subclasses; whereas it can appear in some of them or not under a non homogeneous behavior of the subclasses, leading to a specification or inheritance scenario depending on the significance of W in the superclass. Table $W - S_C$ is an aggregation of all $W - C'$ $\forall C' \in S_C$.

Table 3.7: Inclusion of W as subclasses' descriptor according to diagnostic scenarios.

$(W - S_C)$	***SuperClass***	
Descriptor SubClasses	**Non-Significant**	**Significant**
Non-Significant	None	None
Significant	Some	Some

Finally, Table $W - C - S_C$ (see Table 3.8) is, in fact, the result of aggregating the actions associated to all $W - C - C'$ $\forall C' \in S_C$. The aggregation is $\forall C \in S_C$ in the first row, whereas in the second row is $\exists C' \in S_C$. This shows that, for the first row linked to a $\forall$ operator, the real diagnosis of W cannot be done unless the whole $W - C'$ tables are analyzed together, and leads to introduce a provisional diagnoses for a single C' (see D Table 3.9).

Table 3.8: Synthesis of Table 3.4 and Table 3.7.

$(W - C - S_C)$	***Descriptor SuperClass***			
Descriptor SubClasses	**Non-Significant**		**Significant**	
Non-Significant	$\forall$	None	$\forall$	Only SuperClass
Significant	$\exists$	Some SubClasses	$\exists$	SuperClass and Some SubClasses

Table 3.9: Provisional diagnosis of descriptor W in superclass C and in a subclass C'.

D	***Descriptor SuperClass***	
Descriptor SubClass	**Non-Significant**	**Significant**
Non-Significant	Apparently Irrelevant	Apparently Inconsistence
Significant	Specification	Inheritance

In a second stage, this general framework must be extended to the result of applying Sensitive Analysis, and then Table 3.9 changes to Relations Table ($\mathfrak{R}$) 3.10.

Using the Sensitive Analysis of both partitions $P_\ell, P_{\ell'}$, each descriptor W of a class $C \in P_\ell$ is linked with a term $\mathcal{V}_{W,C}$, and has an associated descriptor-power dp_C. The same happens with subclasses in S_C (getting $dp_{C'}$ $\forall C' \in S_C$). The diagnosis of W regarding C and C' must be done, not considering significance of W in C and C', but considering its descriptor-power dp_C and $dp_{C'}$. These determines a cell on Table 3.10 and provides a provisional diagnoses of W.

The case-analysis of the situations that may occur in a real case correspond to those described in Table 3.10. Table $\mathfrak{R}$ models possible use of W to describe C or $C' \in S_C$ upon its significance in C and in C'. As it happened with D Table (see Table 3.9), the top half are apparent situations (irrelevant or inconsistency) because the final situation is decided by combining diagnoses of all classes (and the aggregation is performed through a $\forall$ operator). Nevertheless, in the bottom half (specifications and inheritances), the aggregation uses the existential operator "exists at least one subclass" ($\exists C' \in S_C$) and might be solved locally as soon as one case goes to one of these cells of the Table $\mathcal{R}$.

Now, there is not anymore a binary scenario and the Sensitive Analysis might produce several evaluations for the subclasses of S_C. I.e, being $S_C = \{C_1, C_2, C_3\}$,
$\mathcal{R}(C, C_1) = Moderate\ Specification$,
$\mathcal{R}(C, C_2) = Weak\ Specification$,
$\mathcal{R}(C, C_1) = Apparently\ Irrelevant$.

According to that, the role of W as a descriptor seems to change from one class to another. This rises the need to introduce an aggregation operator to diagnose the global situation of W regarding the whole subsets S_C. The aggregation is performed now by taking the supreme of the $dp_{C'}$.

A new Table $R - C - S_C$ can be built by the following expression:

$$R - C - S_C(C, S_C) = R - C - C'(C, \psi), \qquad \text{where } \psi = \arg \underset{C' \in S_C}{sup}\ dp_{C'}.$$

In the example, $\psi = C_1$ and the final diagnostic would be that W is in a $Moderate\ Specification$ situation, what means that is not descriptor of C but appears with a Moderate descriptor-power at least in one of the subclasses C' and has no greater discriminant power in any subclass.

The problematic situations for consistency between interpretations of P_ℓ and $P_{\ell'}$ arise for the cells labeled as inconsistency, moderate inconsistency, weak inconsistency or limit inconsistency, as in this situations W seems to be relevant to describe C but disappears from all $C' \in S_C$ in some sense. One of the contributions of this thesis has been to provide tools to reduce inconsistencies as much as possible by proposing new statistical tests to assess significance. Those that persist will be managed at this stage.

Table Actions ($\mathcal{A}$) (see Table 3.11) associates a description activity for each situation described in Table $\mathcal{R}$. In this table, it is indicated the action to be taken in front of each situation described in Table $\mathcal{R}$.

Therefore, a characteristic W will appear or not in the interpretation of the superclass C and in some or not subclasses of S_C according to Table $\mathcal{A}$. Only inconsistencies over certain robustness degree are reported in this description.

The proposed methodology for Consistency Analysis (CNP) is the following:
Given

- W a characteristic
- P_ℓ, $P_{\ell'}$ partitions, where $P_{\ell'}$ is nested in P_ℓ
- $\mathcal{V}_{W,P_\ell}$, $\mathcal{V}_{W,P_{\ell'}}$ descriptors obtained as indicated in Section 3.4

1. Start with maximal descriptors:

    - $\mathcal{V}^{\bullet}_{W,P_\ell} = \mathcal{V}_{W,P_\ell}$

- $\mathcal{V}^{\bullet}_{W,P_{\ell'}} = \mathcal{V}_{W,P_{\ell'}}$

2. $\forall C \in P_\ell$

   (a) Determine $S_C \subseteq P_{\ell'}$.

   (b) Diagnosis:

   - W is irrelevant if:

     $$\forall C' \in S_C, \mathcal{A}(dp_{W,C}, dp_{W,C'}) = None$$

     Thus:

     - $\mathcal{V}^{\bullet}_{W,P_\ell} = \mathcal{V}^{\bullet}_{W,P_\ell} \setminus \{\mathcal{V}_{W,C}\}$
     - $\mathcal{V}^{\bullet}_{W,P_{\ell'}} = \mathcal{V}^{\bullet}_{W,P_{\ell'}} \setminus \{\mathcal{V}_{W,C'}, \forall C' \in S_C\}$

   - W is inconsistent if:

     $$\forall C' \in S_C, \mathcal{A}(dp_{W,C}, dp_{W,C'}) = SupC$$

     Thus:

     - $\mathcal{V}^{\bullet}_{W,P_{\ell'}} = \mathcal{V}^{\bullet}_{W,P_{\ell'}} \setminus \{\mathcal{V}_{W,C'}, \forall C' \in S_C\}$

   - W is Specification if:

     $$\exists C' \in S_C, \mathcal{A}(dp_{W,C}, dp_{W,C'}) = SubC$$

     Thus:

     - $\mathcal{V}^{\bullet}_{W,P_\ell} = \mathcal{V}^{\bullet}_{W,P_\ell} \setminus \{\mathcal{V}_{W,C}\}$
     - $\mathcal{V}^{\bullet}_{W,P_{\ell'}} = \mathcal{V}^{\bullet}_{W,P_{\ell'}} \setminus \{\mathcal{V}_{W,C''}, \forall C'' \in S_C : C' \neq C'' \wedge \mathcal{A}(dp_{W,C}, dp_{W,C''}) = None\}$

   - W is Inheritance if:

     $$\exists C' \in S_C, \mathcal{A}(dp_{W,C}, dp_{W,C'}) = Both$$

     Thus:

     - $\mathcal{V}^{\bullet}_{W,P_{\ell'}} = \mathcal{V}^{\bullet}_{W,P_{\ell'}} \setminus \{\mathcal{V}_{W,C''}, \forall C'' \in S_C : C' \neq C'' \wedge \mathcal{A}(dp_{W,C}, dp_{W,C''}) = SupC\}$

Table 3.10: Relation Table (ℛ): Behaviour of all descriptors using the Sensitive Analysis

	SuperClass					
SubClass	**Robust Non-descriptor**	**Moderate Non-descriptor**	**Weak Non-descriptor**	**Weak Descriptor**	**Moderate Descriptor**	**Robust/Basic Descriptor**
Robust Non-descriptor	Apparently Irrelevant	Apparently Moderate Irrelevant	Apparently Weak Irrelevant	Apparently Weak Inconsistency	Apparently Moderate Inconsistency	Apparently Inconsistency
Moderate Non-descriptor	Apparently Moderate Irrelevant	Apparently Weak Irrelevant	Apparently Limit Irrelevant	Apparently Limit Inconsistency	Apparently Weak Inconsistency	Apparently Moderate Inconsistency
Weak Non-descriptor	Apparently Weak Irrelevant	Apparently Limit Irrelevant	Uncertain	Uncertain	Apparently Limit Inconsistency	Apparently Weak Inconsistency
Weak Descriptor	Weak Specification	Limit Specification	Uncertain	Uncertain	Limit Inheritance	Weak Inheritance
Moderate Descriptor	Moderate Specification	Weak Specification	Limit Specification	Limit Inheritance	Weak Inheritance	Moderate Inheritance
Robust/Basic Descriptor	Specification	Moderate Specification	Weak Specification	Weak Inheritance	Moderate Inheritance	Inheritance

Table 3.11: Action Table ($\mathcal{A}$): Actions of all descriptors using the Sensitive Analysis

	SuperClass					
SubClass	Robust Non-descriptor	Moderate Non-descriptor	Weak Non-descriptor	Weak Descriptor	Moderate Descriptor	Robust/Basic Descriptor
Robust Non-descriptor	None	None	None	None	Inconsistency	Inconsistency
Moderate Non-descriptor	None	None	None	None	Inconsistency	Inconsistency
Weak Non-descriptor	None	None	None	None	Both	Both
Weak Descriptor	None	None	None	None	Both	Both
Moderate Descriptor	SubC	SubC	SubC	SubC	Both	Both
Robust/Basic Descriptor	SubC	SubC	SubC	SubC	Both	Both

3.9 Nested Class Interpretation based on Marginal Significance (NCI-IMS)

A modification of the interpretation methodology CI-IMS is also proposed to guarantee the consistency of an interpretation with future refinements in a new nested partition. Sensitive Analysis methods have been used for this purpose (see Section 3.5).

The methodology CI-IMS described in Section 3.6 is based on the idea of describing the classes according to the attributes giving significant *Test-Value* for the class and the robustness of this significance. Consistency among interpretations of nested partitions requires that significant attributes in the P_ℓ classes must also be significant in some $P_{\ell'}$ classes.

The main objective of the proposed methodology is that the interpretation of P_ℓ (Υ_{P_ℓ}) is consistent with the interpretation of the nested partition $P_{\ell'}$ ($\Upsilon_{P_{\ell'}}$) in such a way that all relevant characteristics of P_ℓ appear as relevant characteristics at least in some subclasses in $P_{\ell'}$. The proposed methodology to interpret nested partitions is the following:

First, both partitions are interpreted with the CI-IMS methodology proposed in Section 3.6. Secondly, both interpretations are analyzed identifying inconsistencies, inheritances, specifications, etc as described in Section CNP 3.8. Then, the interpretations are rebuilt. Given

- $\mathcal{X}$ a set of attributes
- P_ℓ, $P_{\ell'}$ two partitions such that $P_{\ell'}$ is a partition nested in P_ℓ,
- Table Relations ($\mathcal{R}$) 3.10
- Table Actions ($\mathcal{A}$) 3.11.

1. Obtain $\Upsilon_{P_\ell,r}$ the interpretation of P_ℓ using CI-IMS$(\delta, \alpha, \epsilon_1, \epsilon_2, r, \mathcal{X})$ proposed in Section 3.6 with $r = \overline{R}$

2. Obtain $\Upsilon_{P_{\ell'},r}$ the interpretation of $P_{\ell'}$ using CI-IMS$(\delta, \alpha, \epsilon_1, \epsilon_2, r, \mathcal{X})$ proposed in Section 3.6 with $r = \overline{R}$.

3. $\forall W \in \mathcal{D}_W$

   $(\mathcal{V}^\bullet_{W,P^\ell}, \mathcal{V}^\bullet_{W,P^{\ell'}}) = CNP(\mathcal{V}_{W,P^\ell}, \mathcal{V}_{W,P^{\ell'}})$ (see Section 3.8)

4. Build

   $$\Upsilon^\bullet_{P_\ell} = \bigcup_{\forall W \in \mathcal{D}_W} \mathcal{V}^\bullet_{W,P_\ell}$$

   $$\Upsilon^\bullet_{P_{\ell'}} = \bigcup_{\forall W \in \mathcal{D}_W} \mathcal{V}^\bullet_{W,P_{\ell'}}$$

5. $\Upsilon^\bullet_{P_\ell}$ and $\Upsilon^\bullet_{P_{\ell'}}$ are the input for the automatic description using regular expressions as in the methodology CI-IMS explained in Section 3.5.

Examples of superclass YW and subclasses YW-WMBased, YW-WMwSugars, YW-UH

	Classes	**<Woman, Gender>**	**Age**	**C-Reactive Protein**
$\Upsilon^{\bullet}_{YW}$:	YW	$(B, \top)$	$(M, \downarrow)$	$\cancel{(\overline{R}, \times)}$
$\Upsilon^{\bullet}_{YW-WMbased}$:	YW-WMBased	$(B, \top)$	$(W, \downarrow)$	$(M, \uparrow)$
$\Upsilon^{\bullet}_{YW-WMwSugars}$:	YW-WMwSugars	$(B, \top)$	$\cancel{(\overline{R}, \times)}$	$(R, \downarrow)$
$\Upsilon^{\bullet}_{,YW-UH}$:	YW-UH	$(B, \top)$	$(M, \downarrow)$	$\cancel{(\overline{R}, \times)}$

Automatic Descriptions:

YW: The group YW has higher proportion of Women (gender); Age is lower in the group YW.

YW-WMbased: The group YW-WMbased has higher proportion of Women (gender); Age is lower in the group YW-WMbased; C-Reactive Protein is higher in the group YW-WMbased.

YW-WMwSugars The group YW-WMwSugars has higher proportion of Women (gender); C-Reactive Protein is higher in the group YW-WMwSugars.

YW-UH: The group YW-UH has higher proportion of Women (gender); Age is lower in the group YW-UH.

3.10 Characterizing Pre-Post Trajectories

Given two different partitions, the Pre-Post Trajectories analysis allows analyzing how the elements of one partition are distributed in the second partition and vice versa. This can be graphically represented using a bipartite graph where the classes are the nodes and the transitions are the edges between one class of a partition at pre stage to another class of the partition at post stage. In addition, since our problem is an intervention study, this representation can be enriched including the intervention groups.

Thus, there are two partitions - at the beginning (P_o) and at the end of the study (P_f) - and the intervention (T). It is possible to analyze the changes in class membership between pre and post stage depending on the intervention assigned to each person by crossing the initial partition P_o with the intervention T obtaining a new finer partition $P_o \times T$. Therefore, it is expected to see how the participants assigned to certain intervention behave differently from those that are in other intervention groups, even when they have same profiles at the beginning of the study.

Moreover, these resulting trajectories can be used to see the effects of the intervention by projecting the differences in them. These subclasses trajectories can be seen as the product of crossing P_o, intervention T and P_f.

Given $\mathcal{Y}$ a set of attributes, two partitions P_o, P_f and an intervention T,

1. Determine mt as the minimum number of individuals per trajectory to be retained
2. Build $\Psi = \{\phi_1, \ldots, \phi_{n_\Phi}\}$; $\Psi \subseteq P_o \times T \times P_f$; $\phi \in \Psi$ iff $n_\phi \geq mt$ where $n_\phi = card(\phi)$.
3. Build the profile of Ψ using the measures at the beginning of the intervention $\mathcal{Y}_o$ and the NCI-IMS methodology presented in Section 3.9, obtaining $\Upsilon_{\Psi,\mathcal{Y}_o}$.
4. Build the profile of Ψ using the measures at the end of the intervention $\mathcal{Y}_f$ and the NCI-IMS methodology presented in Section 3.9, obtaining $\Upsilon_{\Psi,\mathcal{Y}_f}$.
5. Build the table of profiles comparisons between $\mathcal{Y}_o|\Psi$ versus $\mathcal{Y}_f|\Psi$: $\forall\phi \in \Psi$, compares the profiles $\Upsilon_{\phi,\mathcal{Y}_o}$ and $\Upsilon_{\phi,\mathcal{Y}_f}$ as indicated in Section 3.11. The result identifies significant changes along intervention for all profiles.

3.11 Comparing Two Profiles

Only profiles created with the same attributes are comparable. The following three cases are considered:

1. interpretation of two different classes using the same attributes,
2. interpretation of the same class using same attributes taken under different conditions, for instance, along the time and
3. interpretation of two different classes using different measures of the same attributes taken under different conditions (combinig both points 1 and 2).

In the following explanation, the third case is referred because the other two are particular cases of it.

Given a set of attributes $\mathcal{Y}$ and two classes C, C', the corresponding interpretations can be compared. In the case of pre-post studies, it is interesting to compare the differences of the profiles along the time. Thus, the interpretation of class C is made with the measures of $\mathcal{Y}$ at instant $t-1$ ($\mathcal{Y}_{t-1}$) and the interpretation of C' with the measures at instant t ($\mathcal{Y}_t$) where $t \geq t-1$. However, as it is mentioned, the comparison of two profiles using the same measures or the same classes can be made with the present methodology.

Remember that, a general descriptor $d_{W_C} = (W, C, sense)$ can appear in an interpretation with three possible values in the sense: ↑, -, ↓(see description in Section 3.4).

Comparing the descriptors of both profiles, the following 3 situations can occur (see Table 3.12):

- The descriptor of a characteristic W from an attribute Y does not change between instant $t-1$ and t ($sense(d_{W_{t-1},C}) = sense(d_{W_t,C})$).

- The descriptor of a characteristic W from an attribute Y has change increasing its value between instant $t-1$ and t ($sense(d_{W_{t-1},C}) < sense(d_{W_t,C})$).
- The descriptor of a characteristic W from an attribute Y has change decreasing its value between instant $t-1$ and t ($sense(d_{W_{t-1},C}) > sense(d_{W_t,C})$).

Table 3.12 shows the behavior of operator Φ which models how a descriptor of a characteristic W from an attribute Y changes between $t-1$ and t. For instance, if a characteristic was not descriptor in instant $t-1$ and it becomes a significant descriptor with higher value in instant t, then it means that between both instants the mean or proportion of this characteristic has significantly increased.

Table 3.12: Differences between descriptors in the general case

		descriptor t		
	Φ	-	↑	↓
descriptor t-1	-	-	↑	↓
	↑	↓	-	↓
	↓	↑	↑	-

In this thesis, two methodologies of interpretation have been proposed CI-IMS in Section 3.6 and NCI-IMS in Section 3.9. The resulting descriptors of these methodologies are enriched with the robustness of the descriptor (*descriptor-power*). Thus, Table Φ^* 3.13 is the corresponding extended table of differences of Table Φ 3.12 which takes into account these descriptor-powers.

Table 3.13: Table Φ^*: Differences between descriptors from the CI-IMS or NCI-IMS methodologies.

		Descriptor in instant t													
	Φ^*	$\uparrow \overline{R}$	$\downarrow \overline{R}$	$\uparrow \overline{M}$	$\downarrow \overline{M}$	$\uparrow \overline{W}$	$\downarrow \overline{W}$	↑W	↓W	↑M	↓M	↑R	↓R	$\top$ B	$\perp$ B
Descriptor in instant t-1	$\uparrow \overline{R}$	-	-	-	-	-	-	↑	↓	↑	↓	↑	↓	↑	↓
	$\downarrow \overline{R}$	-	-	-	-	-	-	↑	↓	↑	↓	↑	↓	↑	↓
	$\uparrow \overline{M}$	-	-	-	-	-	-	↑	↓	↑	↓	↑	↓	↑	↓
	$\downarrow \overline{M}$	-	-	-	-	-	-	↑	↓	↑	↓	↑	↓	↑	↓
	$\uparrow \overline{W}$	-	-	-	-	-	-	↑	↓	↑	↓	↑	↓	↑	↓
	$\downarrow \overline{W}$	-	-	-	-	-	-	↑	↓	↑	↓	↑	↓	↑	↓
	↑W	↓	↓	↓	↓	↓	↓	-	↓	-	↓	-	↓	-	↓
	↓W	↑	↑	↑	↑	↑	↑	↑	-	↑	-	↑	-	↑	-
	↑M	↓	↓	↓	↓	↓	↓	-	↓	-	↓	-	↓	-	↓
	↓M	↑	↑	↑	↑	↑	↑	↑	-	↑	-	↑	-	↑	-
	↑R	↓	↓	↓	↓	↓	↓	-	↓	-	↓	-	↓	-	↓
	↓R	↑	↑	↑	↑	↑	↑	↑	-	↑	-	↑	-	↑	-
	$\top$ B	↓	↓	↓	↓	↓	↓	-	↓	-	↓	-	↓	-	↓
	$\perp$ B	↑	↑	↑	↑	↑	↑	↑	-	↑	-	↑	-	↑	-

The difference of two interpretations is assessed as follows.
Given

- $\mathcal{D}_W$ the set of characteristics
- $\Upsilon_C = \{\mathcal{V}_{W,C} : W \in \mathcal{D}_W\}$ the interpretation of class C, where $\mathcal{V}_{W,C} = (W, C, dp_{W,C}, s_{W,C})$ is the descriptor of the characteristic W, $dp_{W,C}$ is the descriptor-power and $s_{W,C}$ the sense.
- $\Upsilon_{C'} = \{\mathcal{V}_{W,C'} : W \in \mathcal{D}_W\}$ the interpretation of class C', $\mathcal{V}_{W,C'} = (W, C', dp_{W,C'}, s_{W,C'})$ is the descriptor of the characteristic W, $dp_{W,C'}$ is the descriptor-power and $s_{W,C'}$ the sense.
- Φ^* the table of differences (see Table 3.13)

Build the Table 3.14:

Table 3.14: Comparison of two profiles

Class	W_1	W_2	...	NumIncreases	NumDecreases
C	$s_{W_1C}\ dp_{W_1C}$	$s_{W_2C}\ dp_{W_2C}$	...	$\underset{s\in\Upsilon_C}{\#}\ s = \uparrow$	$\underset{s\in\Upsilon_C}{\#}\ s = \downarrow$
C'	$s_{W_1C'}\ dp_{W_1C'}$	$s_{W_2C'}\ dp_{W_2C'}$	...	$\underset{s\in\Upsilon_{C'}}{\#}\ s = \uparrow$	$\underset{s\in\Upsilon_{C'}}{\#}\ s = \downarrow$
$\Phi^*{\Upsilon_C\Upsilon_{C'}}$	$\Phi^*(s_{W_1C}, dp_{W_1C}, s_{W_1C'}, dp_{W_1C'})$	$\Phi^*(s_{W_2C}, dp_{W_2C}, s_{W_2C'}, dp_{W_2C'})$	...	$\underset{\Phi^*\in\Phi^*{\Upsilon_C\Upsilon_{C'}}}{\#}\ \Phi^* = \uparrow$	$\underset{\Phi^*\in\Phi^*{\Upsilon_C\Upsilon_{C'}}}{\#}\ \Phi^* = \downarrow$

Then, the final set of differences is: $\Phi^*{\Upsilon_C\Upsilon_{C'}} = \{\Phi^*(s_{W_C}, dp_{W_C}, s_{W'_C}, dp_{W'_C}) : W \in \mathcal{D}_W\}$

The comparison between both profiles can be enriched by adding the knowledge of the semantic of the attributes. For each attribute, it can be considered if it has a direct effect (higher values better) or inverse effect (lower value are better) obtaining *positive* or *negative* effects. For instance, in the case of attributes related with the dietary habits whether it is better to increase or decrease the consumption of a specific food item. For biomarkers, it can be defined whether it is better an increment of its levels or a decrement, etc. Then, two additional columns can be added to Table 3.14:

- Number of increases with positive effect: count of the characteristics with direct effect that have increased.
- Number of decreases with positive effect: count of the characteristics with inverse effect that have decreased.

3.12 Summary

In this chapter, the main contributions of this thesis are presented. First, the formalization of the problem that concerns us is detailed (see Section 3.1). The data of the problem is disaggregated depending on the type of attribute, obtaining a richer model than the traditional one. Then, the methodology that we proposed to analyze this type of problems is presented (see Section 3.2). This methodology is a compendium of different elements that can be used both integrated in the global methodology or independently.

The main idea of the methodology is to cluster the individuals finding their initial and final state. Then, the effects of the intervention are locally analyzed to these obtained groups.

The Integrated Multiview Clustering (IMC) is used to find the different classes both at the beginning and at the end of the study. The profiles of these classes are created using the methodology NCI-IMS because the result of the IMC is a partition nested in other partitions (see Section 3.9). NCI-IMS handles the possible inconsistencies between the profile of a nested class (subclass) and the profile of its corresponding superclass. This methodology uses the resulting profiles of the cluster interpretation methodology CI-IMS (see Section 3.6). CI-IMS is a methodology that, given a partition and a set of attributes, obtains the description of each class according to the significant attributes. This methodology is based on a generalized approach of the *Test-Value* and on the Sensitive Analysis that allows determining the strength of the descriptors.

Finally, a trajectory analysis is performed in order to see the different behavior of the individuals between its initial and final state depending on the assigned intervention (see Section 3.10). The characterization of each resulting trajectory shows for the involved individuals both the degree of adherence to the intervention and the effects of the intervention.

Chapter 4 provides the explanation of the developed tools that are used to carry out this methodology and to display the results. Then, this methodology is applied on our case study in Chapter 5.

Chapter 4

Tools Developed to Support Automatic Reporting Phase

This chapter is divided in three sections. The first section contains the description of the tools that are developed and how to interpret them. T he second section concretes the details about the implementation. Finally, the third section describes a prototype which is the integrated tool that contains all the parts of the proposed methodology in Chapter 3.

Mainly, these tools for representing the results are implemented to support the Cluster Interpretation Methodology. The implementation is in R [R Development Core Team, 2012] (see details in Section 4.2). Using regular expressions, the output of the statistics test and the Class Panel Graphs have been implemented with different formats.

4.1 Description of the Developed Tools

4.1.1 Classical Statistical Tests

The statistical tests are shown in tables. For each attribute, the resulting p-value is shown and a mark if it is significant for a given confidence.

Table 4.1 is an example for Kruskal-Wallis and χ^2-Independence test. The resulting table contains the following 4 columns:

- Pack: Thematic pack of the corresponding attribute
- Attribute: The name of the attribute + explanatory name in brackets
- Type: indicates the type of attribute:
  - N: numerical, Kruskal-Wallis test is applied to this attribute
  - Q: qualitative, χ^2-Independence test is applied to this attribute
  - B: binary, χ^2-Independence test is applied to this attribute

- p-value: The resulting p-value of the test for a given α are the significant ones are marked with "*".

Table 4.1: Example of Kruskal-Wallis test and χ^2 Independence test

Pack	Attribute	Type	p-value
biometrics	sexo (gender)	Q	5.886e-17*
biometrics	EDAD (Age)	N	2.575e-06*
biomarkers	gluc0 (glucose)	N	0.08343
biomarkers	cholest0 (cholesterol)	N	0.2651

N: numerical (Kruskal-Wallis Test was used)
Q: qualitative (χ^2 Independence Test was used)
B: binary (χ^2-Independence Test was used)
*: p-value $< \alpha$

4.1.2 *Test-Value*

The calculation of the *Test-Value* is assessed also by an R script. Different formats of LaTeX tables are implemented. Table 4.2 is an example of a resulting table where the *Test-Value* values are shown with some additional information. In each row are represented the *Test-Value* for an attribute X or for a category s of the attribute X for all the classes in a partition. It contains the following columns:

- Attribute(X)
  - if X is numerical, the name of the attribute
  - if X is qualitative, the name of the attribute and the name of the category s.
- $\overline{X}/P_X$: mean of X/proportion of the category s in X
- Class Mean/Proportion: mean of X^c/proportion of the category s in X^c
- Test-Value: for each class is written the *Test-Value* result and it is marked if it is significant for a given confidence value. The marks of significance are "*" or round by a circle.

Table 4.2: Example of a Test-Value Results with the values of the statistics.

		Class Mean/Proportion			*Test-Value*		
Attribute(X)	$\overline{X}/P_X$	WMb	WMS	UH	t_{WMb}	t_{WMS}	t_{UH}
sexo (gender) - Woman	0.72	0.82	0.74	0.5	2.59*	0.73	-5.15*
EDAD (Age)	44.17	47.11	45.07	37.44	3.03*	1.07	-6.43*
pas_esds_2 (systolic Pressure)	116.33	118.21	115.05	116.44	1.53	-1.19	0.09
fc_d_2 (heart Rate) -	71.35	72.63	71.63	68.78	1.73	0.43	-3.24*

Mean is used for numerical attributes
Proportion is used for binary and qualitative attributes
*: |value|> Numerical:WMb: 2.771 WmS: 2.698 UH: 2.898 Qualitative:2.576

Therefore, each row represents one attribute (numerical) or category (categorical) and each column one class. Note that if the categorical attribute has only two categories, then only one category is represented because the other has the opposite value. Binary attributes are represented as categorical of two categories. The cells corresponding with the statistical value are marked if these are significant depending on a given confidence level. In addition, the sign of the statistic shows whether the mean/proportion of the class is lower or higher respect to the global mean/proportion of the attribute. In the last rows are shown the thresholds to be significants for both numerical (for each class) and qualitative attributes.

The calculation of the *local Test-Value* is assessed by a modification of the previous R script. This new script generates also a LaTeX table but instead of showing the mean/proportion of the global variable is shown the mean/proportion for each superclass.

4.1.3 Cluster Interpretation Methodology

As an intermediate result of the CI-IMS methodology. The Sensitive Analysis is represented by a table such as Table 4.3. This shows the results of the test assessed for each $\nu \in \mathcal{S}$ for an attribute or category and for one class. In this type of tables are shown the name of the attribute, the name of the category if the attribute is categorical, the 5 values of the statistics, the corresponding descriptor power and the sense. The results can be represented by the statistic values, the corresponding p-values or with the symbols ✓and × (three fist lines of Table 4.3). Significance is marked with "*" for the statistic values and p-values and with the symbol ✓. Note that basic descriptors are marked with "B" in the 5 cells because the test is not assessed (see the fourth line of Table 4.3).

		Sample size						
		ν_1	ν_2	ν_3	ν_4	ν_5		
Attribute	**Category**	$n(1-\varepsilon_1-\varepsilon_2)$	$n(1-\varepsilon_1)$	n	$n(1+\varepsilon_1)$	$n(1+\varepsilon_1+\varepsilon_2)$	dp	sense
EDAD (Age)		-1.82	-2.16	-2.58	-2.94*	-3.16*	$\overline{W}$	↓
EDAD (Age)		7.80e-02	4.08e-02	1.54e-02	6.01e-03*	3.14e-03*	$\overline{W}$	↓
EDAD (Age)		×	×	×	✓	✓	$\overline{W}$	↓
sexo (gender)	Woman	B	B	B	B	B	B	⊥
✓: $p - value < \alpha$		test $\in \Theta$						

Table 4.3: Example of Table with the Sensitive Analysis for one Class

Table 4.4 shows an output format of the methodology. In this table the description power and the sense are represented directly for a set of attributes and the classes of a partition. In this table, classes are in the rows and attributes (or categories) in columns. Then, the descriptor power and the sense are indicated in the corresponding cells.

The possible values of the cells are the following:

⊥B : Basic descriptor that the proportion is smaller than a certain boundary (δ) and, in addition, it is weak non-descriptor, moderate non-descriptor or robust non-descriptor.

Table 4.4: Example Table of CI-IMS Methodology

	sexo (gender) - Woman	EDAD (Age)	altura1 (height)	peso1 (weight)	imc1 (BMI)	cint1 (waist)	pas_esds_2 (systolic Pressure)	pad_esds_2 (diastolic Pressure)	fc_d_2 (heart Rate)
M	**↓R**	$\overline{\downarrow \mathrm{W}}$	↑R	↑R	↑R	↑R	↑R	↑M	↓R
YW	**↑R**	↓M	↓R	↓R	↓M	↓R	↓R	↓R	$\overline{\uparrow \mathrm{R}}$
WM	**↑R**	↑R	↓R	↓W	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	↑M	↑R	↑R

⊤B : Basic descriptor that the proportion is higher than a certain boundary $(1 - \delta)$ and it is weak non-descriptor, moderate non-descriptor or robust non-descriptor.

↓ R|M|W : Robust/Moderate/Weak descriptor with lower values than general sample and it is not basic descriptor.

↑ R|M|W : Robust/Moderate/Weak descriptor with higher values than general sample and it is not basic descriptor.

↓R|M|W : Basic descriptor and Robust/Moderate/Weak descriptor with lower values than general sample.

↑R|M|W : Basic descriptor and Robust/Moderate/Weak descriptor with higher values than general sample and.

$\overline{\downarrow \mathrm{W}}|\overline{\uparrow \mathrm{W}}|\overline{\downarrow \mathrm{M}}|\overline{\uparrow \mathrm{M}}|\overline{\downarrow \mathrm{R}}|\overline{\uparrow \mathrm{R}}$: Weak|Moderate|Robust Non-descriptor and no basic descriptor. The arrows represent the sign of the statistic.

Table 4.5 shows the output format of the posterior analysis of nested partitions (methodology NCI-IMS), representing the solution of each situation for an attribute, superclass and subclass. The resolution of each partition is shown in one table, classes are represented in the rows and attributes (or categories) in columns. In this case, only the descriptors that are included in the profile contain values. The value is an arrow indicating the sense of the descriptor. Thus, empty cells represent that the corresponding attribute or category is not included in the corresponding class profile.

Table 4.5: Example table of NCI-IMS Methodology

	p14_1_1 (mainOliveOil) - yes	p14_2_1 (oliveOil) - ≥4spoon	p14_3_1 (vegetables) - ≥2day	p14_4_1 (fruit) - ≥3day	p14_5_1 (redMeat) - ≥1day	p14_6_1 (butter) - ≥1day	p14_7_1 (gasDrinks) - ≥1day	p14_8_1 (wine) - ≥7glass/week	p14_9_1 (legume) - ≥3week	p14_10_1 (fish) - ≥3week	p14_11_1 (commercialBakery) - ≥2week	p14_12_1 (nuts) - ≥3week	p14_13_1 (whiteMeat) - yes	p14_14_1 (sauce) - ≥2week
M	↑		↓		↑	↓				↓	↓		↑	↑
YW	↑				↓	↓		↓	↓				↑	↑
WM	↑				↓	↓	↓	↓	↓				↑	↑

The profiles by class are built using the results of the Cluster Interpretation Methodology both CI-IMS and NCI-IMS.

There are 2 output of the profiles:

- List: For each class, the attribute having significant descriptors are listed indicating whether they have higher or lower significance. The attributes are group by thematic packs in order to make easier the understanding. Additional information such as the descriptor power can be included at the end of each attribute.
- Verbose: For each class, all the significant attributes are described and whether they have a higher or a lower value.

Both for List and Verbose mode, the text is generated using regular expressions. The following list represents the structure of the list. In addition, this list used different colors to mark the robustness of the descriptors.

4.1.3.1 Profile Class *C*

***thematicPack* attributes:**

- ([atName] | [atName] - [cat]) : (Lower|Higher) (R)
- ([atName] | [atName] - [cat]) : (Lower|Higher) (M)

- ([atName] | [atName] - [cat]) : (Lower|Higher) (W)
- [atName] - [cat] : (Lower|Higher) (B)
- **[atName] - [cat] : (Lower|Higher) (BR)**
- **[atName] - [cat] : (Lower|Higher) (BM)**
- **[atName] - [cat] : (Lower|Higher) (BW)**

If the NCI-IMS methodology is used, then the symbols of the items can vary indicating the following:

- Inheritance.

+ Specification.

- Inconsistency.

The *Verbose* mode shows the same information without the list structure. Moreover, the colors are used to mark the robustness of the descriptors and the expression used for each characteristic is the following:

C: [Numerical|Categorical]*

Numerical: [atName] is (lower|higher);
Categorical:(Lower|Higher) proportion of [s] [and (lower|higher) of [s]]* ([atName]);

4.1.4 Class Panel Graphs

Class Panel Graphs is implemented in R and they are also generated in LaTeX code.

An example of Class Panel Graphs where each row represents a class and each column an attribute is in Table 4.6. Also, some variants are implemented:

- Including a row above all the classes representing the distribution of the whole attribute (see Table 4.7).
- Including the results of the interpretation methodology. Then if an attribute is descriptor for certain class, it can be marked in the corresponding cell of the Class Panel Graphs (see Table 4.8). Different symbols can be display for each value of the test. The following symbols are displayed for the proposed methodologies:

  $\perp|\top$: Basic descriptor with lower/higher proportion.

  $\downarrow|\uparrow$: Robust descriptor with lower/higher mean/proportion than general sample and it is not basic descriptor.

↓ R|↑ R : Moderate descriptor with lower/higher mean/proportion than general sample and it is not basic descriptor.

↓|↑ : Weak descriptor with lower/higher mean/proportion than general sample and it is not basic descriptor.

⊥|⊤ : Basic descriptor and Robust descriptor with lower/higher proportion than general sample.

⊥|⊤ : Basic descriptor and Moderate descriptor with lower/higher proportion than general sample.

⊥|⊤ : Basic descriptor and Weak descriptor lower/higher proportion than general sample.

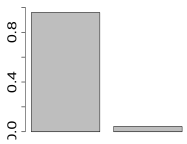
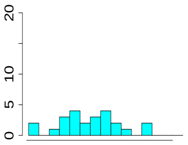
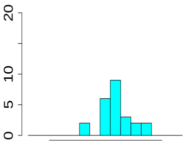
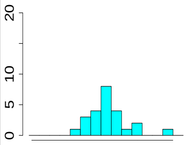
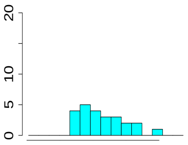
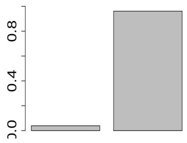
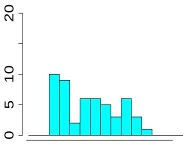
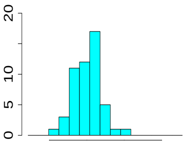
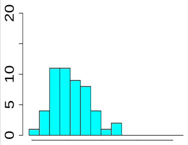
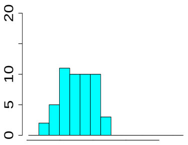
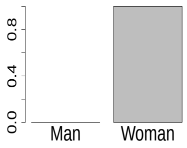
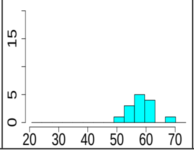
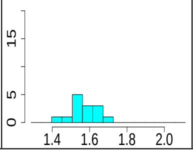
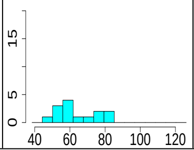
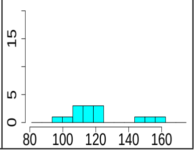
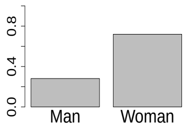
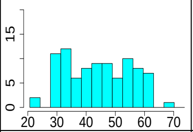
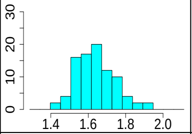
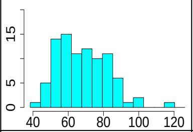
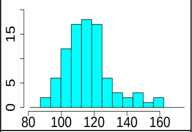
	sexo (gender)	EDAD (Age)	altura1 (height)	peso1 (weight)	pas_esds_2 (systolic Pressure)
M(24)					
YW(51)					
WM(14)					

Table 4.6: Example of Class Panel Graphs

	sexo (gender)	EDAD (Age)	altura1 (height)	peso1 (weight)	pas_esds_2 (systolic Pressure)
All Attribute					

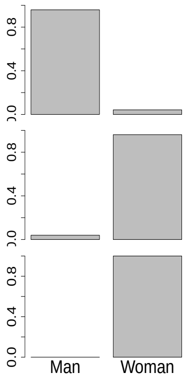
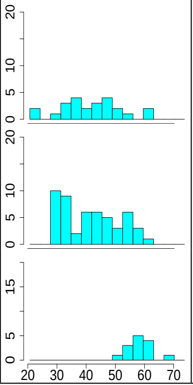
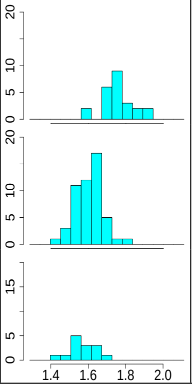
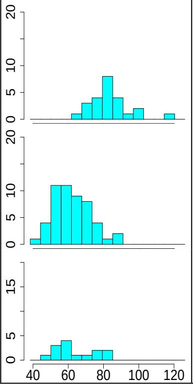
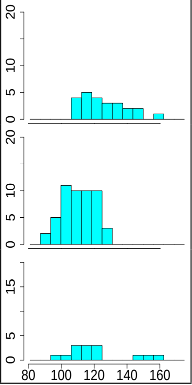
M(24)
YW(51)
WM(14)
					

Table 4.7: Example of Class Panel Graphs with overall distribution in the first row

	sexo (gender)	EDAD (Age)	altura1 (height)	peso1 (weight)	pas_esds_2 (systolic Pressure)

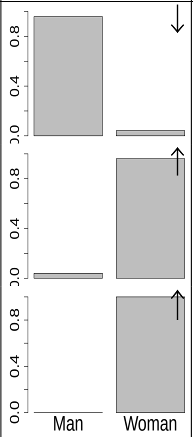
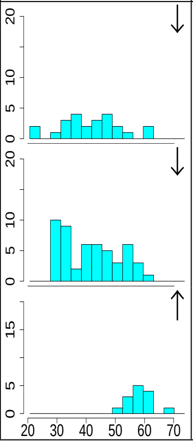
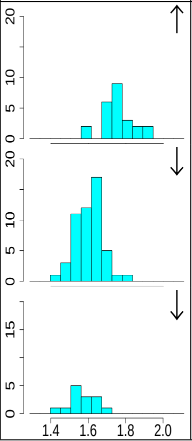
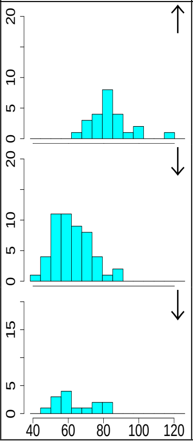
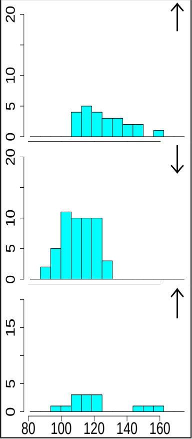
M(24)
YW(51)
WM(14)
					

Table 4.8: Example of Class Panel Graphs with Test-Value

4.1.5 Pre-Post Trajectory Map

Using a bipartite graph the trajectories of the individuals between two or more partitions can be represented.

Since, each partition can be nested into other partitions, this information can be used to build the model. For instance, given two nested partitions at the beginning of the study, the classes of a partition can be represented in nodes, and then, divide each node in its corresponding subclasses. In the case of more nested partitions, this process can be repeated. For the partition at the end of the study is analogous. In addition, since the case study is an intervention study, this representation can be enriched including the intervention groups. The intervention can be represented as other nested partition.

This representation allows graphically identifying how the individuals changes their state in base to the intervention.

Besides, the characterization of each trajectory can be represented by assessing the differences between both profiles, the profile of the initial state and the profile of the final state.

4.1.6 Differences between Profiles

In order to represents the difference between profiles, the following three types of tables are available:

- Table 4.9 shows the items that has increased or decreased from one profile to other. Also, the items are separated according its meaning: healthy items (positive effect) and unhealthy items (negative effect). It is better that the value of a healthy item increases and to decreases in an unhealthy item.

  Table 4.9: Schema of the result table of comparing two profiles

Effect	↑	↓
Healthy Items	Increased	Decreased
Unhealthy Items	Increased	Decreased

- Table 4.10 shows the comparison of the two profiles and the differences. All attributes can be shown or only those changing. Also, the following additional information can be shown:
  - The number of individuals of each intervention.
  - NumIncreases: number of characteristics that increase.
  - NumDecreases: number of characteristics that decrease.

  - Num↑Healthy/Positive: number of characteristics that increase and are healthy/positive.
  - Num↓Unhealthy/Negative: number of characteristics that decrease and are unhealthy/negative.

Table 4.10: Schema of comparing two profiles

	Class	VOO	WOO	Ctrl	p14_1_1 (mainOliveOil) - yes	p14_2_1 (oliveOil) - ≥4spoon	p14_3_1 (vegetables) - ≥2day	p14_4_1 (fruit) - ≥3day	p14_5_1 (redMeat) - ≥1day	p14_6_1 (butter) - ≥1day	p14_7_1 (gasDrinks) - ≥1day	p14_8_1 (wine) - ≥7glass/week	NumIncreases	NumDecreases	Num↑Healthy	Num↓Unhealthy
classO	YW-UH	1	7	1	⊤ B	⊥ B	⊥ B	$\overline{\uparrow \text{R}}$	↑M	↑R	⊥ B	⊥ B	6	7	3	1
classF	W-HOO	4	5	4	⊤ B	⊤ B	$\overline{\uparrow \text{R}}$	$\overline{\uparrow \text{R}}$	⊥ B	⊥ B	↑R	↑W	8	3	7	2
diffs		1	3	0	-	↑	↑	-	↓	↓	↑	↑	6	3	5	3

- Table 4.11 shows the differences from comparing all the profiles at the beginning of the intervention with all the profiles at the end. These comparisons can be filtered by the number of persons that are in the trajectory. In the example, only trajectories between two profiles containing more than 2 individuals are shown. The additional information, as in the previous table, can be shown. Since there are different alternatives from comparing the profiles, the following information is about the number of individuals that are involved in the corresponding comparison[1]:
  - TrajSize: the number of individuals involved in this trajectory.
  - $Size_{C_o}$: the number of individuals used to create the profile of $class_o$
  - $Size_{C_f}$: the number of individuals used to create the profile of $class_f$

[1] In this thesis, usually the comparison of the profiles is made with the individuals that are involved in the trajectory, therefore, $TrajSize = Size_{C_o} = Size_{C_f}$ and only on ot the three is used.

Table 4.11: Scheme of Table comparing all profiles

	$Class_o$	$Class_f$	Int	p14_3_1 - ≥2day	p14_5_1 - ≥1day	p14_6_1 - ≥1day	p14_9_1 - ≥3week	p14_10_1 - ≥3week	p14_11_1 - ≥2week	p14_12_1 - ≥3week	NumIncreases	NumDecreases	Num↑Healthy	Num↓Unhealthy	TrajSize	$Size_{C_o}$	$Size_{C_f}$
1	M-WMbased	M-HOO	VOO	-	-	-	↑	-	-	↓	1	1	0	0	2	2	2
2	M-WMbased	M-HB	VOO	-	-	↑	-	-	-	-	1	0	0	0	2	2	2
3	M-WMbased	M-HB	Ctrl	↑	-	-	-	-	-	-	1	0	1	0	2	2	2
4	M-WMwSugars	M-ProtoCaloric	WOO	-	↑	-	-	-	-	-	1	0	0	0	2	2	2

4.1.7 Graphic of Differences

In order to observe the changes in an attribute at the beginning, and at the end of the study in each class, a "comparative" plots are created. In these graphics, both measures of the attribute are plotted, but conditional to the class. For this representation, the same structure of Class Panel Graphs is used where columns are attributes and rows are classes. Table 4.12 is an example of 4 attributes conditioned to 3 classes and a third categorical attribute is differently shown depending on the type of attribute:

- For numerical attributes, the x-axis represents the values at the beginning of the intervention and the y-axis the values at the end of the intervention. Then, the points can be colored representing the third categorical attribute, that in that case is the assigned intervention.

- For categorical attributes, the bar plot is used. For each category, there two bars showing the proportions at the beginning and at the end of the study. If there is a third categorical attribute defined, then for each category of this third attribute is represented the attribute conditioned to the class and this attribute.

- Binary attributes are treated equal than a categorical attribute with two categories.

The implementation is generic. Thus, the function expects two datasets with the same number of attributes and two optional categorical attributes. Each pair of attributes must be of the same type, and if they are categorical must have the same number of categories to be compared.

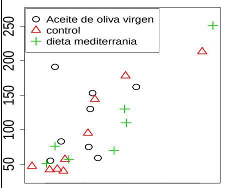
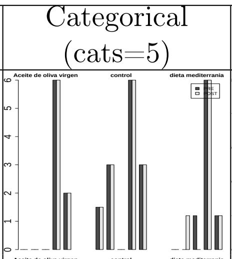
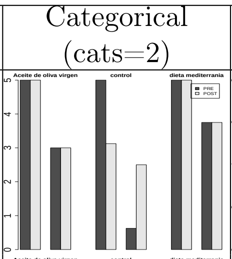
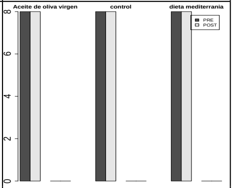
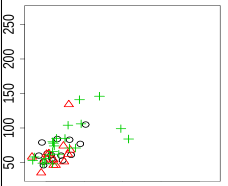
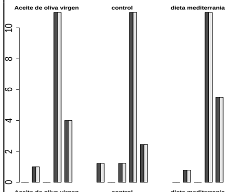
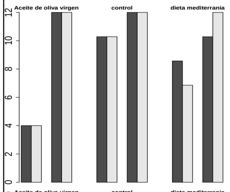
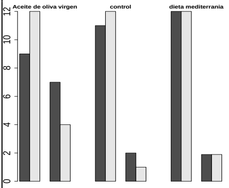
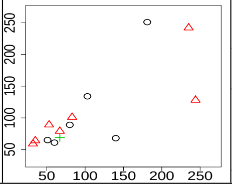
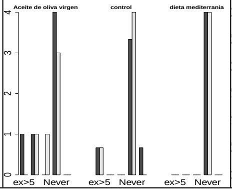
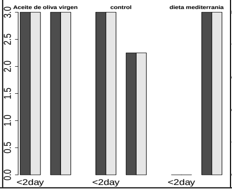
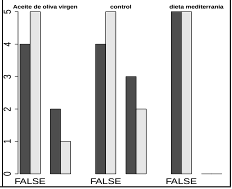
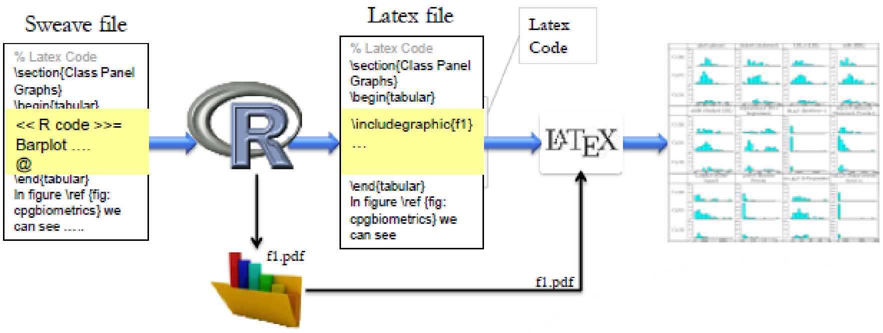
	Numerical	Categorical (cats=5)	Categorical (cats=2)	Binary
M (24)				
WY (51)				
WM (14)				

Table 4.12: Example of Plot3D

4.2 Implementation of the Developed Tools

In this section, the details of the implemented tools are provided. The implementation of the application of this thesis is coded in R [R Development Core Team, 2012]. Also, the *Sweave* package [Leisch, 2002] has been used to create the documentation. Sweave is a tool that allows embedding the R code for complete data analyses in LATEX documents. The purpose is to create dynamic reports. Figure 4.1 is shown the schema of the Sweave framework.



```
% Latex Code
\section{Class Panel
Graphs}
\begin{tabular}
<< R code >>=
Barplot ....
@
\end{tabular}
In figure \ref {fig:
cpgbiometrics} we
can see .....
```

```
% Latex Code
\section{Class Panel
Graphs}
\begin{tabular}
\includegraphic{f1}
...
\end{tabular}
In figure \ref {fig:
cpgbiometrics} we
can see
```

Figure 4.1: Sweave Schema

Therefore, all routines showing results are prepared to create LATEX code.

In order to handle the data, a metadata file was created with the information of all attributes. In this file, for each attribute is indicated:

- English alternative name.
- type (0: invariable in the time, 1: measure before the intervention, 2: measure after the intervention, 3: difference/coefficient).
- Thematic Block: each attribute is assigned to one thematic block for creating the different clusterings.
- the thematic pack for ordering the attributes in the representation of the results.
- role: active, illustrative, dispensable (constants, nulls, ...).
- Original Description of the attribute.
- English Description of the attribute.

With this metadata, it is possible to easily retrieve subsets of attributes in order to build the clustering and to show the results both in thematic blocks and packs.

Both *Class Panel Graphs* and statistical tests are automatically generated with R scripts. These scripts return LaTeX code for a quicker documentation.

4.2.1 Clustering and Cluster Validity Indexes

The clustering is performed using a hierarchical clustering with Ward's method that provides package *stats* included in the core of R [R Development Core Team, 2012]. This function is called with a distance matrix, as it was mentioned. The Gower's distance is implemented in package *StatMatch* [D'Orazio, 2012].

Other clustering methods that are used in this research are built with already existent packages of R: Pam, Kmeans and DBScan from package *fpc* [Hennig, 2013a], Mclust from *mclust* [Fraley and Raftery, 2002].

The cluster validity indexes are assessed with the package *fpc* [Hennig, 2013a].

4.2.2 Statistical Tests

The implementation of the package *stats* is used for both Kruskal-Wallis test and χ^2-Independence test.

The calculation of the *Test-Value* is assessed by an R script implemented from scratch. The following approaches have been implemented:

- The original *Test-Value* defined in Section 2.9.3.

- The modified and generalized variants of *Test-Value* presented in Section 3.3.1 and Section 3.5.
- The local variant of *Test-Value* using nested partitions (see Section 7.1).
- These three variants are implemented in order to handle the dynamic version for comparing the attributes before the intervention against the attributes after the intervention (see Section 3.3.2).

4.2.3 The Cluster Interpretation Methodologies

The CI-IMS methodology was implemented from scratch using the implementations of the Test-Value. The implementation is parametrized, and the most important parameters are the following:

- Partition to be characterized.
- Dataset to be analyzed and an optional second dataset to assess the dynamic Test-Value.
- δ for assessing the basic descriptors.
- ϵ_1 and ϵ_2 to assess the generalized *Test-Value* and determine the descriptors powers.
- Confidence level for assessing the significance whether a descriptor is included in the profile of the class or not.

The NCI-IMS methodology is an extension of the CI-IMS methodology. In that case, both the super partition and the nested partition are needed. From the results of CI-IMS, this methodology uses the specified table of relations in order to rebuild the profiles of all the classes.

The profiles are built using the results of the methodologies. As mentioned, there are 3 output of the profiles: tables, lists and verbose form.

Some table results can be translated to LaTeX using the package *xtable* [Dahl, 2013]. Other tables such the Class Panel Graphs or the table of the *Test-Value* that contains the statistic values have been specially implemented. In order to create the LaTeX tables showing the results of the Cluster Interpretation Methodology, the following latex packages are needed: *rotating*, *tabularx*, *multirow*, *longtable*, *wasyssym* and *testcomp*.

To create the automatic profiles both in List and Verbose mode, the text is generated using regular expressions. These routines need the package *stringr* [Wickham, 2012] to manipulate strings. For compiling the resulting code, the following LaTeX packages are needed in order to correctly print the symbols: *wasyssym* and *testcomp*.

In related research, Section 7.2, an alternative of cluster interpretation using Itemsets has been implemented using the algorithms of association rules of the package *arules* [Hahsler et al., 2012, Hahsler et al., 2005].

4.2.4 Class Panel Graphs

The Class Panel Graphs are implemented in R and also generate LaTeX code. There are two ways of generate them:

- From R: generating a new LaTeX document that contains the CPG of data.
- From Sweave: directly generate the code of the table that contains the CPG.

Given a set of attributes and a partition, the table is created assessing automatically the sizes of the figures to fit in a page, and splitting the table if necessary in several pages. The implementation is highly parametrized: number of columns and rows per page, portrait or landscape mode, the optional test to be shown in the corresponding cells, etc.

The function that creates the Class Panel Graphs needs *lattice* package for creating the breaks of the histograms. Histograms are created with "truehist" from package *MASS* [Venables and Ripley, 2002] because are more parametrized than the default histogram that provides R. Also, package *stringr* [Wickham, 2012] is needed to treat with the class and attribute names in order to automatically assess the width of the columns. For compiling the resulting code, the LaTeX package *rotating* is needed.

4.3 Pre-Post Analysis Tool

The development of an Intelligent Decision Support System that integrates all the parts of the proposed methodology will ease the analysis of Pre-Post studies data.

This interactive tool that automatically generates the Trajectory Map can be useful to see the behavior of the individuals according to the assigned intervention, the profiles of each resulting class and to show per each trajectory the intervention effect.

The idea is that introducing some knowledge, the complete analysis will be assessed as a "black box". Then, the Trajectory Map will be shown. In addition, the information available in this map will be the interpretation of each class and the characterization of each trajectory. Since, usually this type of studies contains large number of attributes, the user will have the option to select those attributes to be shown/compared in each moment.

The previous knowledge that is needed is the following:

- The classification of the attributes in the thematic blocks.
- The classification of the attributes between those that have additive effect or multiplicative effect.
- The intervention and an optional set of attributes that can be used to analyze the adherence to this intervention.

4.4 Summary

The aim of this chapter is to describe the tools that shows some information in order to ease the reading of the results from applying the proposed methodology in our case study in Section 5.

Besides, some specifics about the implementation are given such as the used libraries in Section 4.2.

Finally, the Section 4.3 presents the idea of the final prototype that easily will allow the application of the proposed methodology over the data from any pre-post intervention study.

Chapter 5

Application of the Pre-Post Methodology: a Nutritional Case Study

This chapter presents the application of the proposed methodology in Chapter 3. First, the case study is presented in Section 5.1. This study is a pre-post clinical trial with 3 dietary interventions. At the end of this section, the instantiation of the main steps of the proposed methodology is included.

As in all data mining processes, the first step is the preprocessing of the data. In Section 5.2 the detailed process is described.

The second step is the creation of the thematic blocks (see Section 5.3). These thematic blocks are created from the resulting working data of the preprocessing step.

Then, the IMC is applied over the obtained thematic blocks both for the measures before and after the intervention (see Section 5.4 and Section 5.5). The cluster interpretation methodologies CI-IMS and NCI-IMS are applied over all the resulting partitions and the final profiles are included. The details of both cluster interpretation methodologies are given for two partitions (see Section 5.4.3 and Section 5.4.5).

Once the initial and the final states of the individuals are built, the intervention effects are locally analyzed using both the initial states and the trajectories between the initial and final states of the individuals.

5.1 Description of the Case-Study

The goal of the case study was to analyze the effect of the virgin olive oil within the context of the Mediterranean diet on healthy people. In particular, the aim of the study was to assess whether benefits associated with the traditional Mediterranean diet and virgin olive oil consumption could be mediated through changes in the expression of atherosclerosis-related genes (see [Konstantinidou et al., 2010a, Estruch et al., 2013]).

For this study, a randomized, parallel, controlled clinical trial with 3 dietary interventions was performed. The selected volunteers were randomly assigned to a one of the 3 following dietary interventions:

1. Traditional Mediterranean diet with virgin olive oil (VOO)
2. Traditional Mediterranean diet with washed olive oil (WOO)
3. Control group with participant's habitual diet ($Ctrl$)

First, the 99 volunteers passed an including questionnaire which requested for a preliminary health and a few diet habits. From this first test, 7 people were refused and 2 participants more rejected to participate. In addition, 1 person belonging to the control group declined to participate.

Volunteers were advised by a dietitian to maintain their habitual lifestyle. Exclusion criteria were the following:

- intake of antioxidant supplements;
- intake of acetosalicylic acid or any other drug with established antioxidative properties;
- high levels of physical activity (>3000 kcal/week in leisure-time physical activity);
- obesity [body mass index (BMI) >30 kg/m^2];
- hypercholesterolemia (total cholesterol >8.0 mM or dyslipidemia therapy);
- diabetes (glucose > 126 mg/dl or diabetes treatment);
- hypertension [systolic blood pressure (SBP) >140 mmHg and/or diastolic blood pressure (DBP)>90 mmHg or antihypertensive treatment];
- multiple allergies;
- celiac or other intestinal diseases;
- any condition that could limit the mobility of the subject, making study visits impossible;
- life-threatening illnesses or other diseases or conditions that could worsen adherence to the measurements or treatments;
- vegetarianism or a need for other special diets;
- alcoholism or other drug addiction.

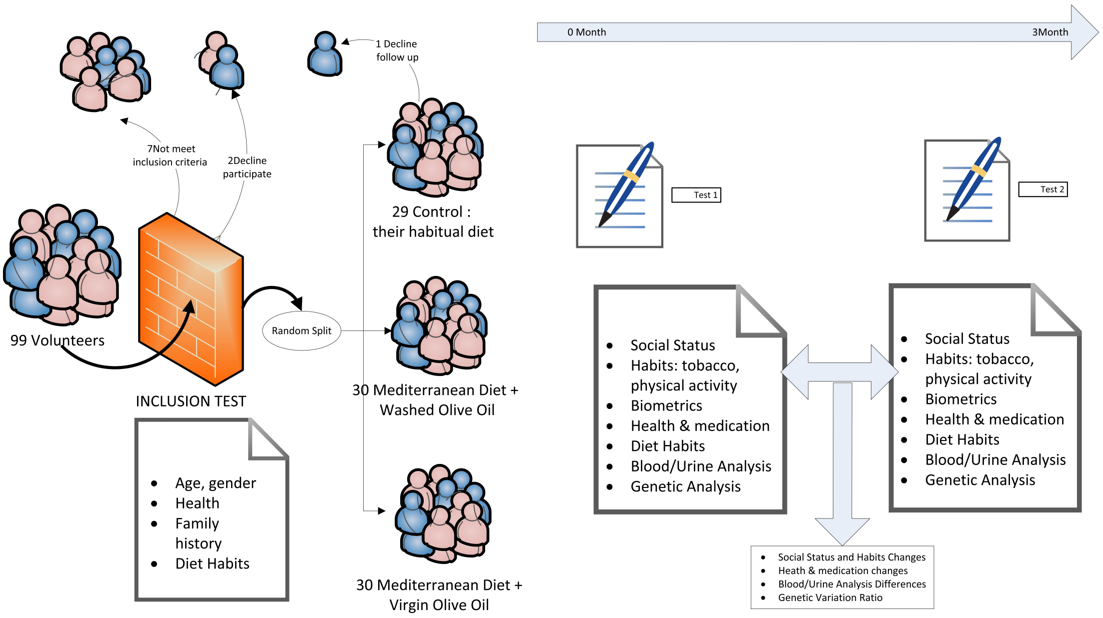


Figure 5.1: Design of the Study

Figure 5.1 depicts the global idea of the study structure. First, an inclusion test defining the subjects that will be in the study. After, subjects were selected and randomly split into the intervention groups. Then, there were two visits where the necessary data was collected: *before* and *after* the diet intervention. Therefore, the study data is collected in 3 phases:

1. Inclusion test
   (a) Biometric characteristics
   (b) Health questionnaire
   (c) Family history questionnaire
   (d) Existing problems about following the study
   (e) Tobacco habits
2. First visit: General questionnaire and First visit tests
   (a) General questionnaire
      i. Biometric characteristics
      ii. Health questionnaire
      iii. Family antecendents questionnaire
      iv. Alcohol habits questionnaire
      v. Sociodemographic questionnaire

   vi. Medication questionnaire

   (b) First Visit tests:

      i. Diet habits questionnaire (p14-item questionnaire [Schröder et al., 2011])
      ii. Physical activity questionnaire
      iii. Blood and Urine analysis
      iv. Genetic analysis

3. Second visit: Following visit after 3 to 9 month since the first visit

   (a) Biometric characteristics
   (b) Changes in sociodemographic conditions
   (c) Health questionnaire
   (d) Tobacco habit questionnaire
   (e) Medication questionnaire
   (f) Diet habits questionnaire (p14-item questionnaire [Schröder et al., 2011])
   (g) Physical activity questionnaire
   (h) Blood and Urine analysis
   (i) Genetic analysis

The inclusion and general questionnaire were complementary. In the second visit, they were requested for some additional health question and the changes of both inclusion and general questionnaires. The diet and physical activity questionnaires, the analysis of blood/urine and genetics tests of the first visit were repeated in the second visit. Besides, the tobacco habits characteristics were collected in the inclusion and second visit.

5.1.1 Sample description

The group of volunteers was divided into 64 women and 25 men between 20 and 64 years old, all holding the exclusion criteria fixed at the experimental design. The vast majority were willing to follow the diet assigned in the study. None of them had any disease where olive oil or dried fruit cannot be ingested. More than 50% declared to have a diet rich in fiber and avoided animal fat.

The volunteers included in the study do not or did not suffer any of the following diseases:

- Myocardial infarction
- Angina
- Embolism or cerebrovascular accident (CVA)

- Arrhythmias or heart disease
- Intermittent claudication
- Diabetes

In addition, 5 subjects had high levels of cholesterol (in 227-280 mg/dL) but none of them took medication for hypercholesterolemia at the beginning of the study.

Only 2 subjects were diagnosed with hypertension and one of them declared to take drugs for this. Although, at the time of the first visit 5 volunteers were taking drugs for hypertension.

5.1.2 Application of the proposed Methodology

The methodology proposed in Section 3.2 is applied over the presented case study. In Figure 5.2 schematically shows the process. The specific steps to apply the methodology are the following:

1. Data Preprocessing (see Section 5.2): cleaning empty and constant attributes, treatment of redundant attributes, null and erroneous values, etc.
2. Create 2 *Thematic Blocks*:
   (a) Divide $\mathcal{X} \cup \mathcal{Y}$ attributes in 2 thematic blocks using the background expertise in the area (see Section 5.3): $\mathcal{B} = \{\mathcal{B}^1(C), \mathcal{B}^2(H)\}$:
   - C contains the baseline characteristics of the individuals.
   - H contains the habits characteristics.

   (b) Build $\mathcal{C}_o$, $\mathcal{H}_o$ and $\mathcal{C}_f$, $\mathcal{H}_f$ accordingly.
3. Profiling the Initial state of individuals (see Section 5.4):
   (a) Build an Integrative Multiview clustering using an agglomerative Hierarchical Clustering with Ward's method and using the Gower's Dissimilarity Coefficient (see section 5.4): obtaining the integrated partition $\mathcal{P}_o = \mathcal{P}_{C_o} \times \mathcal{P}_{H_o}$

   (b) Characterize and interpret the classes of $\mathcal{P}_o$ using $\mathcal{P}_{C_o}, \mathcal{P}_{H_o}$ with the Cluster Interpretation Methodology NCI-IMS (see Section 5.4.5).
4. Profiling the Final state of individuals (see Section 5.5)
   (a) Build an Integrative Multiview clustering: Build the cross clustering using an agglomerative Hierarchical Clustering with Ward's method and using the Gower's Dissimilarity Coefficient (see Section 5.5): obtaining the integrated partition $\mathcal{P}_f = \mathcal{P}_{C_f} \times \mathcal{P}_{H_f}$

   (b) Characterize and interpret the classes of $\mathcal{P}_f$ using $\mathcal{P}_{C_f}, \mathcal{P}_{H_f}$ with the Cluster Interpretation Methodology NCI-IMS (see Section 5.5.6).

5. Trajectory Analysis (see Section 5.6)

6. Analysis of adherence to the Intervention (T) (see Section 5.7)

   (a) The diet habits are directly related with adherence to intervention

   (b) Analysis of the adherence to the intervention over $\mathcal{P}_o$ (see Section 5.7.2).

   (c) Analysis of the adherence to the intervention using the Pre-Post Trajectories (see Section 5.7.3)

7. Analysis of the Intervention effects E using the characterization of the Pre-Post Trajectories Map obtained in Step 5 (see Section 5.8)

   (a) Characterize the additive effect of the intervention depending on each trajectory (see Section 5.8.1)

   (b) Characterize the multiplicative effect of the intervention depending on each trajectory (see Section 5.8.2).

   (c) Build Joint Effect Models (see Section 5.8.3).

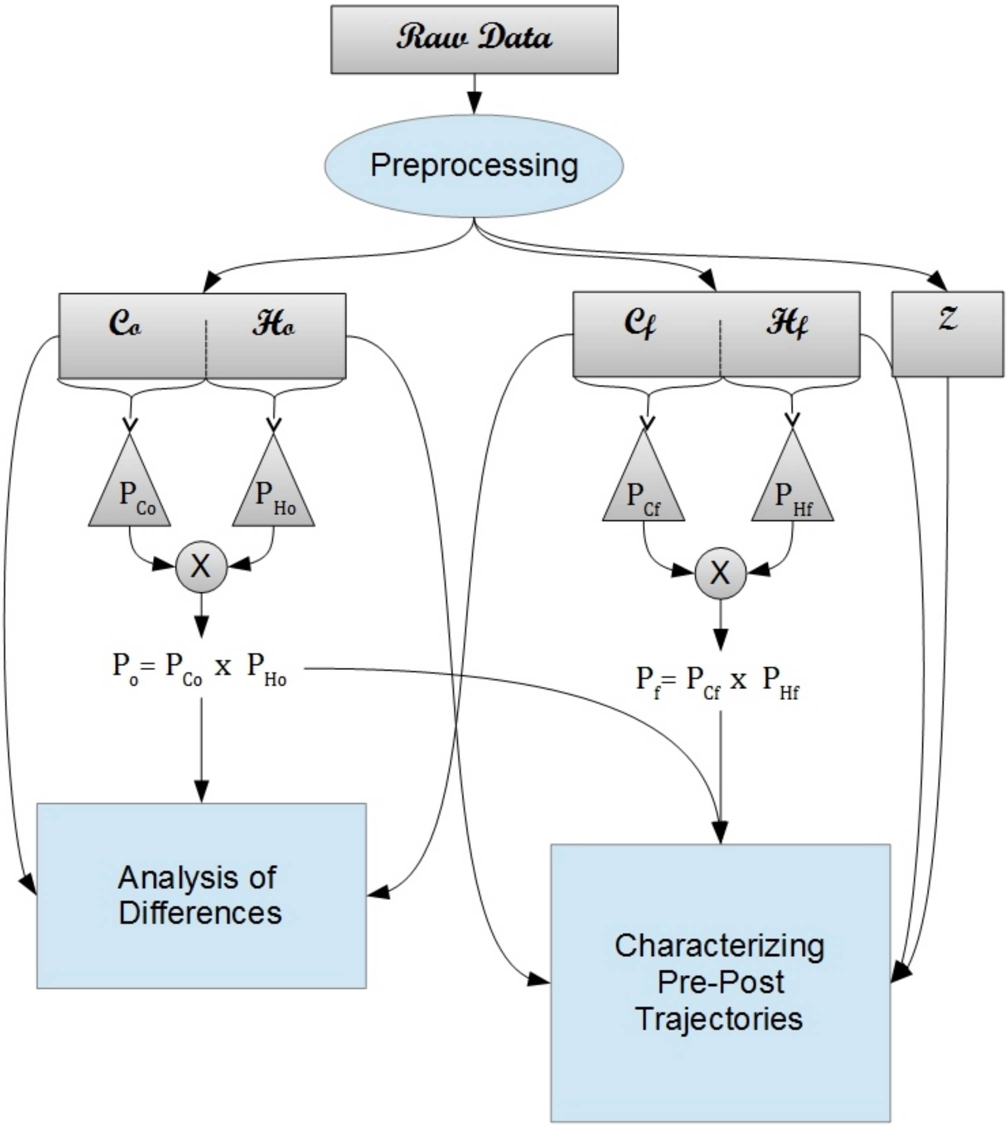


Figure 5.2: Schema of the Applied Methodology over the Case Study

5.2 Preprocessing of Original Data

The original database contained 612 attributes and 89 volunteers/individuals. In the following sections several preprocessing steps were performed to prepare the data for an accurate data mining process.

5.2.1 Cleaning Empty Attributes

The original dataset contained 84 attributes all showing missing values for all individuals. Thus, these attributes were deleted from the dataset because of its lack of information. In the next Table 5.1, these empty attributes are listed. The remaining 528 attributes contained some useful information.

Table 5.1: Empty Attributes

	Name	Description
1	centro	
2	tipo_arr	INCLUSION: Diagnosis arrhythmia
3	año_diag	INCLUSION: Approximate years of diagnosis of diabetes
4	puros	INCLUSION: About how many cigars do you smoke per day?
5	pipas	INCLUSION: About how many pipes do you smoke per day?
6	motiv_ex	INCLUSION: Reason for exclusion
7	excl_mot	INCLUSION: Another reason for exclusion
8	cont_ap1	Surname
9	cont_nom	Name
10	cont_ap2	Second surname
11	con_tel1	Phone 1
12	con_tel2	Phone 2
13	cip	CIP
14	nif	DNI
15	edad_em	GENERAL: Age at diagnosis of pulmonary embolism
16	edad_ane	GENERAL: Age at diagnosis of aortic aneurysm
17	edad_ic	GENERAL: Age at diagnosis of left heart failure
18	edad_tro	GENERAL: Age at diagnosis of deep vein thrombosis
19	edad_ret	GENERAL: Age at diagnosis of retinopathy
20	edad_car	GENERAL: Age ar diagnosis of vascular disease
21	edad_nef	GENERAL: Age at diagnosis of nephropathy
22	eda_epoc	GENERAL: Age at diagnosis of chronic bronchitis - emphysema
23	edad_cat	GENERAL: Age at diagnosis of cataracts
24	edad_apn	GENERAL: Age at diagnosis of sleep apnea
25	edad_dem	GENERAL: Age at diagnosis of dementia
26	edad_par	GENERAL: Age at diagnosis of disease Parckinson
27	med5	GENERAL: Name medication 5.
28	med6	GENERAL: Name medication 6.
29	med8	GENERAL: Name medication 8.
30	med7	GENERAL: Name medication 7.
31	med7a	GENERAL: Medicine 7. morning dosis
32	med7b	GENERAL: Medicine 7. noon dosis
33	med7c	GENERAL: Medicine 7. night dosis
34	med8a	GENERAL: Medicine 8. morning dosis
35	med8b	GENERAL: Medicine 8. noon dosis

36	med8c	GENERAL: Medicine 8. night dosis
37	med5a	GENERAL: Medicine 5. morning dosis
38	med5b	GENERAL: Medicine 5. noon dosis
39	med5c	GENERAL: Medicine 5. night dosis
40	med6a	GENERAL: Medicine 6. morning dosis
41	med6b	GENERAL: Medicine 6. noon dosis
42	med6c	GENERAL: Medicine 6. night dosis
43	pas_esis_1	GENERAL: left upper extremity (sitting patient) 1st taking.PAS
44	pad_esis_1	GENERAL: left upper extremity (sitting patient) 1st taking.PAD
45	fc_a_1	GENERAL: left upper extremity (sitting patient) 1st taking.FC
46	pas_esis_2	GENERAL: left upper extremity (patient seated) 2 taking.PAS
47	pad_esis_2	GENERAL: left upper extremity (patient seated) 2 taking.PAD
48	fc_b_2	GENERAL: left upper extremity (patient seated) 2 taking.FC
49	inciden_2	GENERAL: Incident notes
50	n_trab_c	FOLLOW VISIT 2: What job has the head of household? 3 MONTHS
51	tipoint2	FOLLOW VISIT 2: surgery type 3-MONTHS
52	tipoenf2	FOLLOW VISIT 2: disease type 3-month
53	med24	FOLLOW VISIT 2: Drug Name 4. 3-month
54	med25	FOLLOW VISIT 2: Drug Name 5. 3-month
55	med26	FOLLOW VISIT 2: Drug Name 6. 3-month
56	med28	FOLLOW VISIT 2: Drug Name 7. 3-month
57	med27	FOLLOW VISIT 2: Drug Name 8. 3-month
58	med23a	FOLLOW VISIT 2: Drung 3. morning dosis 3-month
59	med23b	FOLLOW VISIT 2: Drung 3. noon dosis 3-month
60	med23c	FOLLOW VISIT 2: Drung 3. night dosis 3-month
61	med24a	FOLLOW VISIT 2: Drung 4. morning dosis 3-month
62	med24b	FOLLOW VISIT 2: Drung 4. noon dosis 3-month
63	med24c	FOLLOW VISIT 2: Drung 4. night dosis 3-month
64	med25a	FOLLOW VISIT 2: Drung 5. morning dosis 3-month
65	med25b	FOLLOW VISIT 2: Drung 5. noon dosis 3-month
66	med25c	FOLLOW VISIT 2: Drung 5. night dosis 3-month
67	med26a	FOLLOW VISIT 2: Drung 6. morning dosis 3-month
68	med26b	FOLLOW VISIT 2: Drung 6. noon dosis 3-month
69	med26c	FOLLOW VISIT 2: Drung 6. night dosis 3-month
70	med27a	FOLLOW VISIT 2: Drung 7. morning dosis 3-month
71	med27b	FOLLOW VISIT 2: Drung 7. noon dosis 3-month
72	med27c	FOLLOW VISIT 2: Drung 7. night dosis 3-month
73	med28a	FOLLOW VISIT 2: Drung 8. morning dosis 3-month
74	med28b	FOLLOW VISIT 2: Drung 8. noon dosis 3-month
75	med28c	FOLLOW VISIT 2: Drung 8. night dosis 3-month
76	inciden2	FOLLOW VISIT 2: Incindent notes 3-month
77	nos2aantes	NOS2A measeured expression in RQ before treatment (baseline)
78	nos2adespues	NOS2A measeured expression after treatment in RQ (3weeks)
79	nox1ante	NOX1 measeured expression in RQ before treatment (baseline)
80	nox1despues	NOX1 measeured expression after treatment in RQ (3weeks)
81	ratio_rq_nos2a	RATIO_RQ_NOS2A after/before
82	ratio_rq_nox1	RATIO_RQ_NOX1 after/ants
83	log2ratiorq_nos2a	log2ratioRQ_NOS2A
84	log2ratiorq_nox1	log2ratioRQ_NOX1

5.2.2 Cleaning Constant Attributes

There were 58 attributes containing a single constant value for all individuals. This kind of attributes showed common characteristics of the volunteers, useful to describe the sample characteristics. Having a constant value, these were not discriminant and were not included into the analysis.

The remaining 470 attributes were informative without constant values. These 58 attributes are enumerated Table 5.2.

Table 5.2: Constant Attributes

Name	Constant Value	Description
nodo	40	Node
tipoparticip	0	Participant Type
Seguimiento	0	
proceden	europa	INC: Origin
cambdon	NO	INCLUSION: Do you think moving to another town in the next few years or have a limitation that prevents or hinders controls and attend scheduled meetings?
prob_die	NO	INCLUSION: Do you informed by health workers suffering from an illness that prevents you follow any particular diet that includes olive oil and / or nuts it has been?
iam_0	NO	INCLUSION: Do you informed by health personnel who have ever had a myocardial infarction has been?
angor_0	NO	INCLUSION: Do you informed by medical personnel ever had angina has been?
avc_0	NO	INCLUSION: Do you informed by health personnel who have ever had a stroke or a stroke has been?
arritmia	NO	INCLUSION: Do you informed by medical personnel ever had any heart disease or arrhythmias has been?
claudica	NO	INCLUSION: Do you informed by medical personnel ever had has been intermittent claudication?
diabetes	NO	INCLUSION: Do you informed by health workers who have had diabetes has been?
tra_col	NO	INCLUSION: Do you follow any Hypolipidemic agents treatment?
inclus	SI	Inclusion: Inclusion
embolia	NO	GENERAL: Diagnosed pulmonary embolism?
aneuris	NO	GENERAL: Diagnosed aortic aneurysm?
icizq	NO	GENERAL: Diagnosed left heart failure?
troboli	NO	GENERAL: Diagnosed deep vein thrombosis?
bronqui	NO	GENERAL: Diagnosed chronic bronchitis - emphysema?
catarata	NO	GENERAL: Diagnosed Cataracts?
apneas	NO	GENERAL: Diagnosed sleep apnea?
demenc	NO	GENERAL: Diagnosed dementia?
parckin	NO	GENERAL: Diagnosed with Parckinson disease?
retino	NO	GENERAL: Diagnosed Retinopathy?
cardio	NO	GENERAL: Diagnosed vascular disease?
nefro	NO	GENERAL: Diagnosed nephropathy?
mol_bebe	NO	GENERAL: Has it ever bothered you people criticizing your drinking?
beb_mal	NO	GENERAL: Have you ever felt bad or guilty about your drinking?
beb_meno	NO	GENERAL: Have you had the impression that you should drink less?
beb_mañ	NO	GENERAL: Have you ever drank at first thing in the morning to steady your nerves or to get rid of a hangover?

cardiov1	NO	GENERAL: In the past month have you taken heart medication?
hipocol1	NO	GENERAL: In the past month have you taken cholesterol medication?
insulin1	NO	GENERAL: In the past month have you taken insulin?
ado1	NO	GENERAL: In the past month have you taken medication for diabetes? (Other than insulin)
visita_1_14p	0	P14 VISIT 1
visita_1_af	0	
seg	3	Follow (in months)
iam2	NO	FOLLOW VISIT 2: Have you been informed by health personnel of having a heart attack in the past year? 3 MONTHS
angor2	NO	FOLLOW VISIT 2: Have you been told by medical personnel of having angina in the last year? 3 MONTHS
avc2	NO	FOLLOW VISIT 2: Have you been informed by health personnel of suffering a stroke in the past year has been? 3 MONTHS
diabete2	NO	FOLLOW VISIT 2: Have you been diagnosed in the last year of diabetes? 3 MONTHS
hipecol2	NO	FOLLOW VISIT 2: Have you been informed by health personnel of having high cholesterol? 3 MONTHS
angiop2	NO	FOLLOW VISIT 2: Have you been informed by health personnel of having a coronary angioplasty with or without stenting or Coron bypass surgery
paro2	NO	FOLLOW VISIT 2: Have you been informed by health personnelof suffering a heart attack from which you have recovered in the last year? 3 MONTHS
aneuri2	NO	FOLLOW VISIT 2: Have you been diagnosed in the last year of aortic aneurysm? 3 MONTHS
claudi2	NO	FOLLOW VISIT 2: In the last year, have been diagnosed with a circulatory disorder in the legs? 3 MONTHS
iqvasc2	NO	FOLLOW VISIT 2: Has this circulatory disorder been the cause of an intervention or amputation of part of a limb? 3 MONTHS
arritm2	NO	FOLLOW VISIT 2: Do you informed by health personnel of suffering an arrhythmia in the last year? 3 MONTHS
nefrodb2	NO	FOLLOW VISIT 2: In diabetic patients, have you been diagnosed in the last year some of the following complications: Renal affectation? 3-month
dialis2	NO	FOLLOW VISIT 2: In case of renal affectation, the deterioration of renal function has caused entering a dialysis program? 3 MONTHS
retinop2	NO	FOLLOW VISIT 2: Affectation of the retina from diabetes (diabetic Retinopaía) that triggered a laser treatment 3-month
catarat2	NO	FOLLOW VISIT 2: Have you been diagnosed in the last year of cataracts? 3 MONTHS
intquir2	NO	FOLLOW VISIT 2: Have you been surgically intervented this past year? 3 MONTHS
enferm2	NO	FOLLOW VISIT 2: Have you developed in the last year some kind of disease that you have not previously diagnosed (if yes indicate which) 3-MONTHS?
insulin2	NO	FOLLOW VISIT 2: During the past month have you taken insulin? 3 MONTHS
ado2	NO	FOLLOW VISIT 2: During the past month have you taken medication for diabetes? (Other than insulin) 3 MONTHS
visita_2_14p	3	VISIT P14 2
visita_2_af	3	

Therefore, from these constant attributes, the sample can be defined as follows: All volunteers were European. None of them had limitations to follow all visits until the end of the study. No subject took drugs for hypercholesterolemia. Volunteers had no problems with the alcohol consumption. They did not take heart drugs, insulin nor other diabetes medication.

None subject had reported to have high levels of cholesterol in the second visit.

The list of diseases which none of the volunteers suffered included in Table 5.2 are the following: myocardial infarction, angina, embolism or cerebrovascular accident(CVA), Arrhythmias or heart disease, intermittent claudication, diabetes, pulmonary embolism, aneurysm left heart failure, deep vein thrombosis, chronic bronchitis - emphysema, cataracts, sleep, apnea, dementia, Parkinson, retinopathy, heart-vascular disease, nephropathy, coronary angioplasty, heart attack/failure, renal condition with Dialysis treatment, surgical intervention, diabetic retinopathy and no diseases were detected during the study.

This configures a sample of volunteers that can follow a diet intervention without risks.

5.2.3 Cleaning Drug Attributes

The pharmacological treatments followed by every person were requested in the 1st and 2nd visit questionnaire. Information of 11 different families of drugs was asked and the person can provide up to 8 different specific drug names with corresponding doses. This implies a total of 43 attributes in the database with information on medication before the study (1st visit) and other 43 after the study (2nd visit). An additional attribute (*cam_med* (change_medication)) reported changes on treatments during the study.

The attributes defining the medication treatments in the 1st visit questionnaire were the following:

- 11 attributes asking whether the person uses drugs of some family or not.
  - 7 attributes with useful information:

    aspirin1 GENERAL: In the past month have you taken aspirin, Adiro or similar?

    aines1 GENERAL: In the past month have you taken other drugs to relieve pain or fever?

    tranqui1 GENERAL: In the past month have you taken tranquilizers, sedatives, anxiety pills or sleeping pills?

    vitamin1 GENERAL: In the past month have you taken vitamin or mineral?

    hipoten1 GENERAL: In the past month have you taken heart medication?

    hormo1 GENERAL: In the past month have you taken hormone therapy? (Only women)

    otromed1 GENERAL: In the past month have you taken any other medicines?
  - 4 attributes are constants and as it said in Section 5.2.2, these will be not included in the analysis.

    cardiov1 GENERAL: In the past month have you taken heart medication?

    hipocol1 GENERAL: In the past month have you taken cholesterol medication?

    insulin1 GENERAL: In the past month have you taken insulin?

    ado1 GENERAL: In the past month have you taken medication for diabetes? (Other than insulin)
- 8 attributes to introduce specific drug names
  - 4 attributes with useful information

  - 4 attributes with all missing values (cleaned in Section 5.2.1)

- 8*3 attributes to introduce the doses of the specific drugs in the morning, noon, night:
  - 12 attributes with a few values
  - 12 attributes with all missing values (cleaned in Section 5.2.1)

For the 2nd visit questionnaire, the 44 attributes are the following:

- 11 attributes asking whether the person uses drugs of some family or not.
  - 9 attributes with useful information:

    aspirin2 FOLLOW VISIT 2: During the past month have you taken aspirin, Adiro or similar? 3 MONTHS

    aines2 FOLLOW VISIT 2: During the past month have you taken other drugs to relieve pain or fever? 3 MONTHS

    tranqu2 FOLLOW VISIT 2: During the past month have you taken tranquilizers, sedatives, anxiety pills or sleeping pills? 3 MONTHS

    vitamin2 FOLLOW VISIT 2: During the past month have you taken vitamin or mineral? 3 MONTHS

    cardiov2 FOLLOW VISIT 2: During the past month have you taken heart medication? 3 MONTHS

    hipoten2 FOLLOW VISIT 2: During the past month have you taken medication for high blood pressure? 3 MONTHS

    hipocol2 FOLLOW VISIT 2: During the past month have you taken cholesterol medication? 3 MONTHS *(one single no missing value)*

    hormo2 FOLLOW VISIT 2: During the past month have you taken hormone therapy? (Women only) 3 MONTHS

    otromed2 FOLLOW VISIT 2: During the past month have you taken any other medicines? 3 MONTHS
  - 2 attributes are constants and as it is said in Section 5.2.2, these will be not included in the analysis

    insulin2 FOLLOW VISIT 2: During the past month have you taken insulin? 3 MONTHS

    ado2 FOLLOW VISIT 2: During the past month have you taken medication for diabetes? (Other than insulin) 3 MONTHS

- 8 attributes to introduce specific drug names
  - 3 attributes with useful values
  - 5 attributes with all missing values (cleaned in Section 5.2.1)

- 8*3 attributes to introduce the doses of specific drugs in morning, noon, night:
  - 6 attributes with few values
  - 18 attributes with all missing values (cleaned in Section 5.2.1)

- 1 question asking for changes (yes/no)

Notwithstanding, some inconsistencies were observed in this part of the questionnaire. Some subjects declared "no" or nothing (missing) taking drugs of a certain family while they reported drugs of that family in the list of drugs (see Table 5.3 and Table 5.4).

Figure 5.3 shows the content of the 4 specific drugs attributes that contained useful information. Only one person used 4 different drugs. Only 3 persons used 3 different drugs. Only 15 persons used 2 different drugs. Table 5.3 shows the frequency table for these drugs. About the half of the sample took some medication, but no particular drugs were consumed more than others. In fact, none of the drugs is taken by more than 3 persons.

Figure 5.4 and Table 5.4 show the same information related to the second visit. In that case, only 2 persons took more than 2 drugs and it can be seen that some of the persons ended part of their treatments along the study.

Table 5.3: Drugs frequencies for the 1st visit

	med1	**med2**	**med3**	**med4**	**Total**
ACTONEL	1	0	0	0	1
ADOFE	0	1	0	0	1
AIRTAL	2	0	0	0	2
ALPRANOLAN	0	1	0	0	1
AMOXICILINA	1	0	0	0	1
ANTALGIN S P	0	0	1	0	1
APROVEL	1	0	0	0	1
ASPIRINA	0	1	0	0	1
ASPIRINA SP	2	0	0	0	2
ATENOLOL	1	0	0	0	1
ATENOLOL 50	0	1	0	0	1
AUXINA E	0	1	0	0	1
BISOLGRIP	0	1	0	0	1
BOLTIN	1	0	0	0	1
CALCIO D	0	1	0	0	1
CROSINOR	1	0	0	0	1
DIAZEPAM S P	0	1	0	0	1
DIAZEPAN 5	1	0	0	0	1
DOLMEN	1	0	0	0	1
EBASTEL S P	2	0	0	0	2
ESPERCORBIN	1	0	0	0	1
ESPORADIC	1	0	0	0	1
GELOCATIL S P	2	0	0	0	2
GELOCATIL SP	1	0	0	0	1
HIDROFEROL	0	0	1	0	1
IBUPROFENO	1	0	0	0	1
IBUPROFENO S P	1	0	0	0	1
IBUPROFENO S P	3	0	0	0	3
LEXATIN	1	0	0	0	1
LISINOPRINA	0	1	0	0	1
LORATADINA	0	0	1	0	1
LORMETAZEPAN 1	1	0	0	0	1

MYOLATAN	0	1	0	0	1
NOCTAMIT	1	0	0	0	1
NORVAS 10 S P	1	0	0	0	1
OMEPRAZOL S P	1	0	0	0	1
OPENVAS 40	1	0	0	0	1
OSVICAL	0	1	0	0	1
PARACETAMOL	0	1	0	0	1
PARACETAMOL 1	1	0	0	0	1
PARACETAMOL S B	0	1	0	0	1
PARACETAMOL S P	2	0	0	0	2
PARCHES EURA	0	1	0	0	1
PSEOXETINA	1	0	0	0	1
SALIDUR 1 3D	1	0	0	0	1
SIBELIUM	1	0	0	0	1
TERMALGIN	1	0	0	0	1
VENTOLIN S P	1	1	0	0	2
YASMIN	1	0	0	0	1
ZILDRIC 150	0	0	0	1	1
ZOMIG 1 MES	1	0	0	0	1
Total	40	15	3	1	59

Table 5.4: Drugs frequencies for the 2$^{\text{nd}}$ visit

	med21	**med22**	**med23**	**Total**
1	0	0	1	1
AAS S P	1	0	0	1
ACTONEL	1	0	0	1
AIRTAL	1	0	0	1
ALPRANOLAM	0	1	0	1
ATENOLOL	1	0	0	1
BOLTIN	1	0	0	1
CALCIO D	0	1	0	1
DIAZEPAN	1	0	0	1
DIAZEPAN S P	0	1	0	1
DOLMEN S P	1	0	0	1
DORMIDINA	1	0	0	1
EBASTEL S P	1	0	0	1
ESPORADIL	1	0	0	1
HIDROFEROL	0	0	1	1
IBUPROFENO	1	0	0	1
IBUPROFENO S P	4	0	0	4
JIBÉLIUM	1	0	0	1
LEXATIN	1	0	0	1
LISINOPRINA	0	1	0	1
LORACEPAN	0	1	0	1
LORMATAZEPAN 1	1	0	0	1
ONEPRAZOL S P	1	0	0	1
PARACETAMOL	1	0	0	1
PARACETAMOL S P	1	0	0	1
PARCHES EURA	0	1	0	1
PAROXETINA	1	0	0	1
POLSRSMIOE	1	0	0	1
SALIDUR 1 3D	1	0	0	1

SCETILCISTEINA	1	0	0	1
SIMUASTATINA	1	0	0	1
SINTABAC	1	0	0	1
TERMALGIN S D	1	0	0	1
TERUASMIN	1	0	0	1
VENTOLIN S P	1	1	0	2
VOLTIN	1	0	0	1
YASMIN	1	0	0	1
ZOMIG	1	0	0	1
Total	33	7	2	42

In Table 5.5 a list with all drugs reported by subjects is built. The drugs are classified in the 11 drug families. New binary attributes recording whether a person used drugs for a certain family or not were built and were used instead of the original attributes containing inconsistencies.

The missing values in drug names were considered as negative answer unless the person reported taking drugs of that family. This means a "Yes" on the final drug family attribute, even though the specific drug name was not provided.

Let F_D be the new attribute to be created corresponding to some drug family (D=$\{$*Aspirin, Antiinflammatory, Anxiolytic, VitaMin, Heart, Hypertension, Hypercholesterol, Insulin, Diabetes, Hormon, Other*$\}$). Let D_1, D_2 ... D_{f_D} the set of attributes reporting specific drugs taken by users for family F_D. Let f_D be the binary original attribute for family D

$$F_D = \begin{cases} 1 & \text{if } f_D = 1 \\ 1 & \text{if } D_1 \vee D_2 \vee ... \vee D_{f_D} = 1 \\ 0 & otherwise \end{cases} \tag{5.1}$$

Table 5.5: Specific drug names classified by families

Drug Family	**Specific Drug Names**		
Aspirin	• Aspirina	• Ass SP	• Dolmen
Antiinflammatory	• Airtal • Antalgin • Hespercorbin	• Gelocatil • Ibuprofeno	• Paracetamol • Termalgin
Anxiolytic	• Adofen • Alpranolan • Diazepam • Dormidina	• Esporadil • Frosinor • Lexatin • Lorazepam	• Lormetazepam • Noctamit • Paroxetina
Vitamin	• Auxina E • Calcio D	• Hidroferol • Osvical	
Heart	*no drugs reported*		

Hypertension	• Aprovel • Atenolol • Lisinoprina • Norvas • Openvas • Salidur
Hypercholesterol	• Simuastatina
Insulin	*no drugs reported*
Diabetes	*no drugs reported*
Hormon	• Boltin • Parches Eura • Yasmin
Other	• 1 • Actonel • Amoxicilina • Bisolgrip • Ebastel • Loratadina • Myolastan • Omeoprazol • Polaramine • Scetilcisteina • Sibelium • Sildric/ Zildric • Sintabac • Teruasmin • Ventolin • Zomig

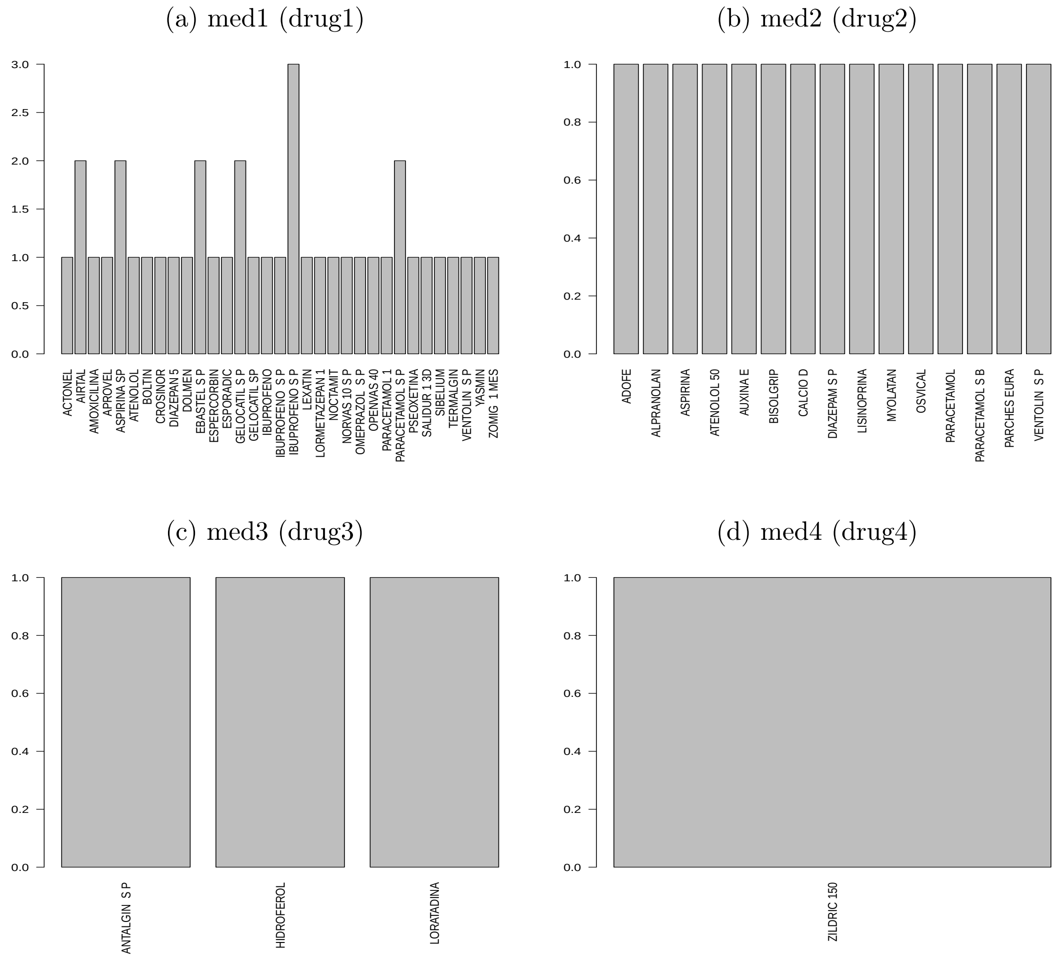
The 11 families were recoded for both the 1st and 2nd visit, following the same procedure. For these 11 new attributes in the 1st visit, the attributes *Heart_1, Insulin_1, Hypercholesterol_1* and *Diabetes_1* are constant and will be not used in the analysis; but instead, they will be used for the sample description. For 2nd visit, 1 person was taking some drugs for the *Hypercholesterol_2* but *Heart_2, Insulin_2, Diabetes_2* were also constant and the last three are not used in the analysis.

Also, the global indicator: *cam_med* (changes in medication) has been recomputed according to cleaned indicators.

In addition, participants were requested to fill the doses but just a few persons did it. That is the reason to drop these attributes off the analysis.

Finally, the medication treatments were defined by 16 attributes: 7 attributes of the 1st visit, 8 attributes of the 2nd visit and 1 additional attribute for the changes in medication during the study.

Therefore, the initial 87 attributes for medication treatments in the database were replaced by 16 attributes. Then, 444 attributes contained useful information.

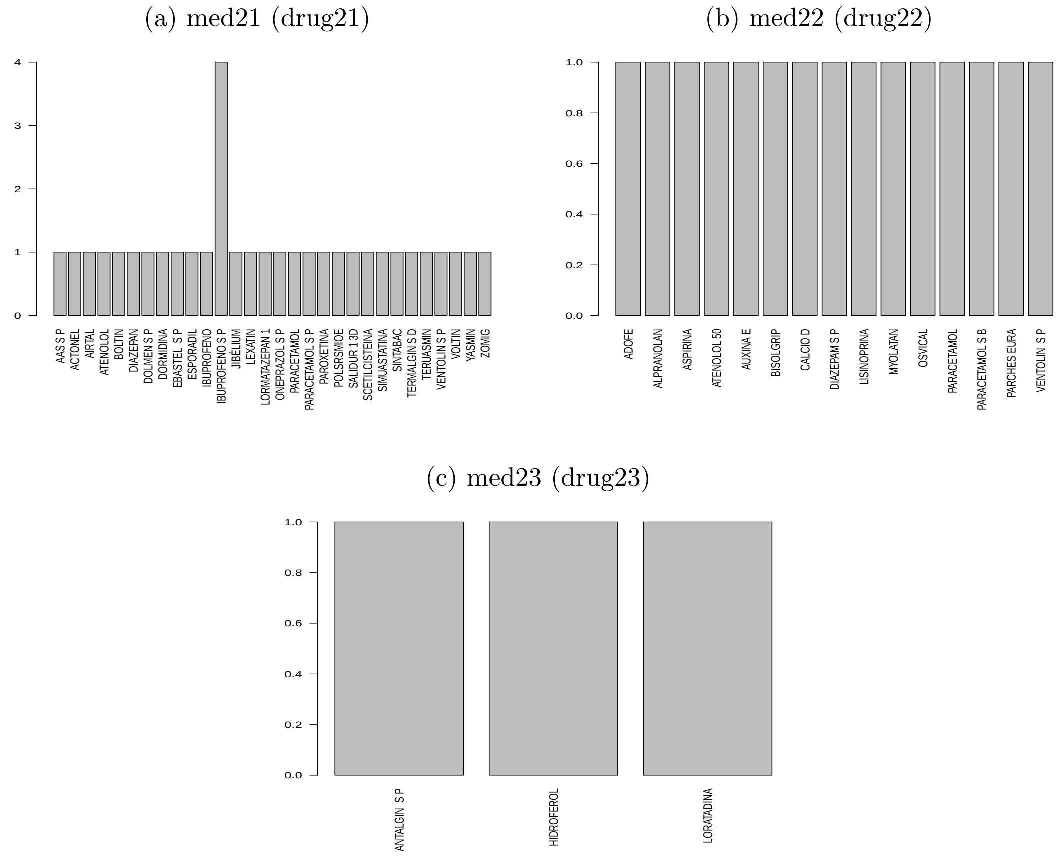

(a) med1 (drug1)

(b) med2 (drug2)

(c) med3 (drug3)

(d) med4 (drug4)

Figure 5.3: Specific drugs for the 1st visit



(a) med21 (drug21)

(b) med22 (drug22)

(c) med23 (drug23)

Figure 5.4: Specific drugs for the 2$^{\text{st}}$ visit

5.2.4 Treatment of Redundant Attributes and Recategorizations

The dataset contained 37 attributes that were redundant and only one per group was retained for the analysis since it did not give additional information.

These attribute groups are the following:

- 3 attributes describing volunteer.
  - *id* (id)
  - *paciente* (id)
  - *VAR00004* (id)

id (id) was used.

- 4 attributes describing the dietary intervention group of each volunteer.
  - *grup_int* (diet Group)
  - *grupo* (dietGroup)
  - *VAR00001* (dietGroup)
  - *VAR00002* (diet Group)

grup_int (diet Group) was used.

- 2 attributes being the weight of the volunteer. The first attribute was requested in the inclusion test and the second was measured in the first visit.
  - *peso_inc* (weight)
  - *peso1* (weight)

The measured attribute *peso1* (weight) was used.

- 2 attributes being the height of the volunteer. The first attribute was requested in the inclusion test and the second was measured in the first visit.
  - *altura* (height)
  - *altura1* (height)

The measured attribute *altura1* (height) was used.

- 2 attributes being the BMI of the volunteer. The first attribute was requested in the inclusion test and the second was measured in the first visit.
  - *imc_inc* (BMI)
  - *imc1* (BMI)

The measured attribute *imc1* (BMI) was used.

- 3 biometric attributes referring to the systolic pressure. 1 attribute was requested in the inclusion test and the other 2 were taken in the first visit.
  - *psist0* (sistolic Pressure)
  - *pas_ esds_ 1* (systolic Pressure)
  - *pas_ esds_ 2* (systolic Pressure)

The second taking *pas_ esds_ 2* (systolic Pressure) was used.

- 3 biometric attributes referring to the diastolic pressure. 1 attribute was requested in the inclusion test and the other 2 were taken in the first visit.
  - *pdias0* (distolic Pressure)
  - *pad_ esds_ 1* (diastolic Pressure)
  - *pad_ esds_ 2* (diastolic Pressure)

The second taking *pad_ esds_ 2* (diastolic Pressure) was used.

- 2 biometric attributes referring to the heart rate. Both attributes corresponded to two blood pressure taken in the first visit.
  - *fc_ c_ 2* (heart Rate)
  - *fc_ d_ 2* (heart Rate)

The second taking *fc_ d_ 2* (heart Rate) was used.

- 2 biomarkers measuring the cholesterol. The first attribute was requested in the inclusion and the second was measured in the first visit.
  - *col_ inc* (cholesterol)
  - *cholest0* (cholesterol)

The measured attribute *cholest0* (cholesterol) was used.

- 2 biomarkers measuring the HDL. The first attribute was requested in the inclusion and the second was measured in the first visit.
  - *hdl_ inc* (HDL)
  - *hdl0* (HDL)

The measured attributes *hdl0* (HDL) was used.

- 2 biomarkers measuring the LDL. The first attribute was requested in the inclusion and the second was measured in the first visit.
  - *ldl_inc*
  - *LDL_0* (LDL)

The measured attribute *LDL_0* (LDL) was used.

- 2 biomarkers measuring the triglycerides. The first attribute is requested in the inclusion and the second is measured in the first visit.
  - *trigl_in* (triglycerides)
  - *tryg0* (triglycerides)

The measured attribute *tryg0* (triglycerides) was used.

- 2 attributes measuring the tyrosol in urine in the 1$^{\text{st}}$ test.
  - *tyru0* (Tyrosol)
  - *tyr0* (tyrosol)

tyru0 (Tyrosol) was used.

- 2 attributes measuring the tyrosol in urine in the 2$^{\text{nd}}$ test.
  - *tyru3* (Tyrosol)
  - *tyr3* (tyrosol)

tyru3 (Tyrosol) was used.

- 2 attributes measuring the OH−tyrosol in urine in the 1$^{\text{st}}$ test.
  - *ohtyru0* (OHTyrosol)
  - *ohtyr0* (ohtyrosol)

ohtyru0 (OHTyrosol) was used.

- 2 attributes measuring the OH−tyrosol in urine in the 2$^{\text{nd}}$ test.
  - *ohtyru3* (OHTyrosol)
  - *ohtyr3* (ohtyrosol)

ohtyru3 (OHTyrosol) was used.

Also, there were a couple of cases in which the original attributes had been either discretized or categorized. As a general principle, the original attributes were maintained in the analysis.

- 1 attribute discretizing the age (*EDAD*) of the volunteers
  - *CAedad_44* (age categorized in two groups)

EDAD (Age) was used.

- 2 attributes recategorizing the *grup_int* (diet Group) in a binary indicator of control/diet intervention group.
  - *grup_dm* (dietGroupMed)
  - *groupos_2* (dietGroupMed)

grup_int (diet Group) was maintained.

5.2.5 Treatment of Null Values

The dataset contained some null values having specific meaning, and proper imputation was inferred from other attributes values. In next Table 5.6, the values that were detected to be null for each attribute are shown, and in the last column, the "action" performed is specified.

Table 5.6: Null Values per attribute

Name	**Description**	**Null**	**Action**
Cigarril	INCLUSION: how many cigarettes do you smoke per day?	88, MISSING	if tabaco≠Yes , Change to 0
edad_men	GENERAL: Only women: When your menopause started?	88, 99	Create a binary indicator *menopausia_bin* (menopause) = $\begin{cases} Yes, \ if \ edad\_men \ has \ a \ value; \\ No, \ if \ edad\_men = 88, 89 \end{cases}$
est_civ2	FOLLOW VISIT 2: Civil status 3-Month	MISSING	if cam_est = no, copy(est_civ1)
n_perho	FOLLOW VISIT 2: Number of people with whom they share home 3-month	MISSING	if cam_pho = No, copy(numper)
cam_est	FOLLOW VISIT 2: Has your civil status changed since your last visit? 3 MONTHS	MISSING	if est_civi = est_civ2 , No else Yes
cam_pho	FOLLOW VISIT 2: Have you changed the number of people with whom they share the home since the last visit? 3 MONTHS	MISSING	if numper=n_perho, No else Yes
n_treball	FOLLOW VISIT 2: What concrete work do or did? 3 MONTHS	MISSING	if cam_slab=No, copy(trab_pac)
sit_labo	FOLLOW VISIT 2: What is your current employment status? 3 MONTHS	MISSING	if cam_slab=No, copy(sitlabor)
cam_slab	FOLLOW VISIT 2: Have you changed your employment status since the last visit? 3 MONTHS	MISSING	if trab_pac=n_treball & sitlabor=sit_labo, No else Yes
cigarr2	FOLLOW VISIT 2: How many cigarettes do you smoke per day? 3 MONTHS	88	if tabaco≠Yes, Change 0

puros2	FOLLOW VISIT 2: How many cigars do you smoke per day? 3 MONTHS	88, 85	if tabaco≠Yes, Change 0
pipas2	FOLLOW VISIT 2: How many pipes do you smoke per day? 3 MONTHS	88	if tabaco≠Yes, Change 0
cam_tab	FOLLOW VISIT 2: Have you changed your smoking habits in the last 6 months?	MISSING	if tabaco= tabaco2 & cigarril=cigarr2 & puros=puros2 & pipas=pipas2 then No; else Yes
tabaco2	FOLLOW VISIT 2: Do you currently smoke cigarettes? 3 MONTHS	MISSING	if cam_tab = No, copy(tabaco)

5.2.6 Treatment of Erroneous Values

The database contained some inconsistencies that were identified as erroneous values by the experts. Some of them were inferred from other attributes, but not all. Table 5.7 corresponds to the inconsistencies in several attributes: column "New Value" shows the replaced value and column "Reason" details whether the solution came from the experts or it was inferred from other attributes.

Table 5.7: Erroneous values per attribute

Name	ID	Value	New Value	Reason
pertensa	400086	67	Change to 6.5	after consulting the experts
il_6_0	400078	-50	Change to 0	after consulting the experts
il_8_0	400078	-20	Change to 0	after consulting the experts
cam_est	400088	sí	Change to no	est_civ = est_civ2
cam_est	400013	no	Change to yes	est_civ ≠ est_civ2
cam_slab	400004	no	Change to yes	sitlabor ≠ sit_labo

In next Table 5.8, inconsistencies of subjects answering the attributes related to the tobacco habit were collected. In this table, the column "Action" shows how the inconsistencies were replaced with new values.

Table 5.8: Inconsistencies in tobacco habit attributes

ID	tabaco	años_tab	cigarril	tabaco2	cigarr2	puros2	pipas2	cam_tab	Action
400069	Yes	40	25	Yes	25	MISSING	MISSING	yes*	cam_tab=No
400047	ex>5*	88	88	ex1a5*	88	88	88	no	??
400063	Yes	20	20	Yes	8	MISSING	MISSING	no*	cam_tab=Yes
400070	MISSING*	15	15	Yes	15	MISSING	MISSING	no	tabaco=Yes
400030	ex1a5*	88	88	ex0a1*	MISSING	MISSING	MISSING	no	??
400054	Yes	MISSING*	1	Yes	1	MISSING	MISSING	no	años_tab=MISSING
400068	Yes	23	5	Never*	5	MISSING	MISSING	no	tabaco2=Yes
400085	Yes	9	2	Yes	8	MISSING	MISSING	no*	cam_tab=Yes
400004	Yes	10	20	Yes	20	MISSING	MISSING	yes*	cam_tab=No

Finally, there was a set of attributes computing the pre-post differences of several parameters. Mostly biometric characteristics or biomarkers of the individuals measured in both 1st and 2nd visit. Self computed difference will be used in this study instead of those provided in the database, because it contained some inconsistencies and did not include all repeated measures.

Given the small concentrations registered by the genetic attributes, the effect of the intervention was computed under a multiplicative assumption rather than additive. Therefore, instead of differences, ratios were computed, both in original and logarithmic scale: ratio (post/pre) and $log_2(ratio)$.

To this point, 458 attributes were considered including the new pre-post difference attributes.

5.2.7 Treatment of irrelevant attributes

The dataset contained several attributes that were irrelevant for our purposes. Therefore, they were not included in the analysis. In the next list, they are enumerated.

- id (identifier)
- 39 attributes related to dates:

- *dianac* (birth Day)
- *mesnac* (birth Month)
- *añonac* (birth Year)
- *diaexam0* (inc Exam Day)
- *mesexam0* (inc Exam Month)
- *añoexam0* (inc Exam Year)
- *fechaexam0* (inc Exam Date)
- *diaexam1* (gralVisitDay)
- *mesexam1* (gralVisitMonth)
- *añoexam1* (gralVisitYear)
- *fechaexam1* (gralVisitDate)
- *acei_ di1* (lastOliveOilDay)
- *acei_ me1* (lastOliveOilMonth)
- *vino_ di1* (lastWineDay)
- *vino_ me1* (lastWineMonth)
- *acei_ añ1* (lastOliveOilYear)
- *vino_ añ1* (lastWineYear)
- *diap14_ 1* (firstVisitDay)
- *mesp14_ 1* (firstVisitMonth)
- *anop14_ 1* (firstVisitYear)
- *dia_ af_ 1* (firstAfVisitDay)
- *mes_ af_ 1* (firstAfVisitMonth)
- *ano_ af_ 1* (firstAfVisitYear)
- *diaexam2* (secondVisitDay)
- *mesexam2* (secondVisitMonth)
- *añoexam2* (secondVisitYear)
- *fechaexam2* (secondVisitDate)
- *acei_ di2* (lastOliveOilDay)
- *acei_ me2* (lastOliveOilMonth)
- *acei_ añ2* (lastOliveOilYear)
- *vino_ di2* (lastWineDay)
- *vino_ me2* (lastWineMonth)
- *vino_ añ2* (lastWineYear)
- *diap14_ 2* (secondp14VisitDay)
- *mesp14_ 2* (secondp14Visitmonth)
- *anop14_ 2* (secondp14VisitYear)
- *dia_ af_ 2* (secondAfVisitDay)
- *mes_ af_ 2* (secondAfVisitMonth)
- *ano_ af_ 2* (secondAfVisityear)

5.2.8 Final Working Data Matrix

The resulting dataset contains 89 subjects and 418 attributes.

In Section 5.1, a first profile of the volunteers configured with inclusion/exclusion criteria has been introduced and in Section 5.2.2, has been extended with the constant attributes characterizing the sample. After analyzing the rest of attributes, the complete profile is detailed.

The group of volunteers was divided into 64 women and 25 men between 20 and 64 years old. The vast majority were willing to follow the diet assigned in the study and none of them had a physical limitation which made the study visits impossible. None of them had any disease where olive oil or dried fruit cannot be ingested. More than 50% declared to have a diet rich in fiber and to avoid animal fat.

Basically, volunteers were Spanish (86 subjects), 1 German, 1 Italian and 1 from Venezuela. Besides, almost all Spaniards were from Catalonia. 70% of the sample was married, 22% single and the rest were in other situations. At the beginning of the study, only 1 person was retired and the rest were actively working. At the end, two subjects changed their work status from working to housewife and retired.

Volunteers had no problems with the alcohol consumption; over 20% of subjects were smokers, around 12% were ex-smokers and the rest had never smoked.

None of the volunteers suffered any of the following diseases before nor during the study: Myocardial infarction, Angina, Embolism or cerebrovascular accident(CVA), Arrhytmias or heart disease, Intermitent claudication, Diabetes, Pulmonary embolism, Aneurysm, Left heart failure (insuficiencia cardiaca izquierda), Deep vein thrombosis, Chronic bronchitis - emphysema, Cataracts, Sleep apnea, Dementia, Parkinson, Retinopathy, Heart-vascular disease, Nephropathy, Coronary angioplasty, Heart attack, Renal condition with Dialysis treatment, Surgical intervention, Diabetic retinopathy or hypercholesterolemia (except for one individual that started medication for that during the study).

Only one person had depression and took medication for that. 5 women had cancer in the past - 4 mama and 1 uterine fibroid. 4 volunteers had suffered a bone fracture in the past. 16 women (25%) had menopause.

Half of the sample took some drugs but none of them took drugs for heart or diabetes. None of them was taking more than 4 drugs. During the study, 22 subjects had changed their initial medication.

13 subjects had relatives that have died because of myocardial infarction or angina.

The participants were described by 418 attributes which can be group by their thematic into packs. Table 5.9 contains the attributes of each thematic pack. Note, that this grouping into thematic packs has only representative purpose.

Table 5.9: Attribute list classified by thematic packs

Thematic Pack	**Attribute List**
Diet Intervention	grup_int (dietGroup)
Biometric	sexo (gender), EDAD (Age), altura1** (height), peso1** (weight), imc1** (CMI), cint1** (waist), pas_esds_2*** (sistolicPressure), pad_esds_2*** (distolicPressure), fc_d_2*** (heartRate)
Socio-demographic	escolar (estudies), paisnac (Country), lugarnac (birthPlace), est_civi** (CivilState), numper** (numPerson), sitlabor** (workStatus), trab_pac* (work)
tobacco habits	tabaco** (tobacco), cigarril** (numCigarettes), aqos_tab(yearsSmoking), puros2** (cigar), pipas2** (pipe)
diet habits	c_grasan (avoidAnimalFat), c_fruta (richFiber), acei_ho1* (oliveOilToday), vino_ho1* (wineToday), p14_1_1** (mainOliveOil), p14_2_1** (oliveOil), p14_3_1** (vegetables), p14_4_1** (fruit), p14_5_1** (redMeat), p14_6_1** (butter), p14_7_1** (gasDrinks), p14_8_1** (wine), p14_9_1** (legumes), p14_10_1** (fish), p14_11_1** (commercialBakery), p14_12_1** (nuts), p14_13_1** (whiteMeat), p14_14_1** (sauce), cap_cam (adaptDiet)
physical activity habits	ge2_3_5s_1** (lightWeek), ge4_5_5s_1** (moderateWeek), ge6_12s_1** (intenseWeek), gehos_1** (homeWorkWeek), getots_1** (totalWeek), ge2_3_5a_1** (lightYear), ge4_5_5a_1** (moderateYear), ge6_12a_1** (intenseYear), gehoa_1** (homeWorkYear), getota_1** (totalYear)
Menopause & Diseases	hipercol, hta0*, trathta0, pertensa (stressful), depre (hasDepression), cancer (hasCancer), fractura (hasBoneFracture), tipo_can (cancerType), edad_fra (fractureAge), edad_dep (DepressionAge), edad_can (cancerAge), edad_men, menopausia_bin (hasMenopause), disnea** (hasDyspnea
Medication	maspirina1** (medAspirin), mdolor1** (medNSAID), mansiedad1** (medAnxiolytic), mvitMin1** (medVinMin), mtension1** (medHeart), mhormo1** (medHormon), motros1** (medOther), cam_med (change_medication)
Familiar antecedents	ant_iam (famMyocardialInfartion), fam_col (famColestherol) , fam_hta (famBloodPressure) , fam_can (famCancer), fam_exit (famHeartDiseases), fam_avc(cerebrovascularAccident), trab_cab (famWork)
Biomarkers	oxldl0** (OxidizedLDL), hdl0** (HDL), gluc0** (glucose), cholest0** (cholesterol), pcr0** (CReactiveProtein), tryg0** (triglycerides), tyru0** (Tyrosol), ohtyru0** (OHTyrosol), mohtyu0** (MOH_tyrosol), il_6_0** (IL6), il_8_0** (IL8), il_10_0** (IL10), mcp_1_0** (MCP1), t_pa_0** (tPA), ifn_g_0** (IFNg), s_cd40l_0** (sCD40L), tnf_a_0** (TNFa), s_vcam_1_0, s_p_selectin_0** (sPselectin), oxo_gg_0** (8_OXOgG), isoprostanos_0** (F2αIsoprostanes), LDL_0** (LDL)
Genetic	abca1antes**, abcg1antes**, adam17ante, adamts1antes**, adrb2antes**, aldh1a1antes**, anxa1antes**, arhgap15antes**, arhgap19antes**, arhgef6antes**, ccng1antes**, cd36antes**, cetpantes**, chukantes**, dclre1cantes**, ercc5antes**, ifna1antes**, ifngantes**, il10antes**, il6antes**, il7rantes**, liasantes**, mpoantes**, msr1antes**, nfkb2antes**, nr1h2antes**, nr1h3antes**, ogtantes**, olr1antes**, osbpantes**, pla2g4bantes**, polkantes**, pparaantes**, pparbpantes**, ppardantes**, ppargantes**, ptgs1antes**, ptgs2antes**, rgs2antes**, scarb1antes**, tnfsf10antes**, tnfsf12_tnfsf13antes**, tp53antes**, usp48antes**, xrcc5antes**
Unclassified	filter_$, outliers

* Means that the attribute has a repeated measure after the intervention
** Means that the attribute has a repeated measure at the end of the study and the difference pre-post has been computed
*** Means that the attribute has not a repeated measure at the end of the study but the difference pre-post is in the dataset

5.3 Creation of the Thematic Blocks

Clustering these subjects is complicated due to the fact that the big amount of attributes includes a wide range of aspects. For this reason, it is proposed to split these attributes into different blocks and then, to cluster the subjects in such a way that each clustering tackles one attribute block with own meaning.

From the original dataset of 612 attributes, 238 attributes not providing useful information for clustering were cleaned in the preprocessing (see Section 5.2), and 15 attributes were derived from recoding the date of menopause and drugs attributes. All differences were computed and replaced those that existed. 180 attributes are related to gene expression. The final 418 attributes contains:

- 10 intrinsic attributes (including the identifier and intervention)
- 76 initial attributes
- 76 final attributes
- 180 attribute referring to gene expression
- 76 difference attributes

From the remaining attributes, 134 have been selected for creating the states of the individuals: 8 intrinsic and 126 that corresponds to 63 initial measures and 63 final measures. 14 attributes were discarded because do not provide useful information for out purpose.

The thematic block $C = \mathcal{B}^1$ is composed by the selection of 47 attributes that show the initial characteristics (baseline) of the individuals. Table 5.10 lists these attributes. In this table, the attributes are grouped depending on the sub-thematic. In column "Intrinsic" are marked those that are constant during the study and in column "Type" is shown whether the attribute is numerical (N), qualitative (Q) or binary (B).

Pack	Attribute	Intrinsic	Type
biometrics	sexo (gender)	yes	Q
biometrics	EDAD (Age)	yes	N
biometrics	altura1 (height)	yes	N
biometrics	peso1 (weight)	no	N
biometrics	imc1 (BMI)	no	N
biometrics	cint1 (waist)	no	N
biometrics	pas_esds_2 (systolic Pressure)	no	N
biometrics	pad_esds_2 (diastolic Pressure)	no	N
biometrics	fc_d_2 (heart Rate)	no	N
tobacco	tabaco (tobacco)	no	Q
sociodemographic	est_civi (civilState)	no	Q
sociodemographic	trab_pac (work)	no	Q
diseases	menopausia_bin (menopause)	yes	B
diseases	pertensa (stressful)	yes	N
N:numerical Q: qualitative B:binary			

Pack	Attribute	Intrinsic	Type
diseases	depre (depression)	yes	B
diseases	cancer (cancer)	yes	B
diseases	fractura (bone fracture)	yes	B
diseases	disnea (dyspnea)	no	B
drugs	maspirina1 (aspirin)	no	B
drugs	mdolor1 (NSAID)	no	B
drugs	mansiedad1 (anxiolytic)	no	B
drugs	mvitMin1 (vitamin or minerals)	no	B
drugs	mtension1 (tension)	no	B
drugs	mhormo1 (hormones)	no	B
drugs	motros1 (other)	no	B
biomarkers	gluc0 (glucose)	no	N
biomarkers	cholest0 (cholesterol)	no	N
biomarkers	LDL_0 (LDL)	no	N
biomarkers	hdl0 (HDL)	no	N
biomarkers	tryg0 (triglycerides)	no	N
biomarkers	oxldl0 (Oxidized LDL)	no	N
biomarkers	isoprostanos_0 (F2 α Isoprostanes)	no	N
biomarkers	ifn_g_0 (Interferon-γ)	no	N
biomarkers	mcp_1_0 (Monocyte Chemotactic Protein-1)	no	N
biomarkers	s_p_selectin_0 (sP-selectin)	no	N
biomarkers	s_cd40l_0 (sCD40 Ligand)	no	N
biomarkers	pcr0 (C-Reactive Protein)	no	N
biomarkers	oxo_gg_0 (8-Oxoguanine)	no	N
biomarkers	tnf_a_0 (Tumor necrosis factor-α)	no	N
biomarkers	t_pa_0 (Tissue Plasminogen Activator)	no	N
biomarkers	s_vcam_1_0 (Soluble cell adhesion molecules-1)	no	N
biomarkers	il_6_0 (Interleukin 6)	no	N
biomarkers	il_8_0 (Interleukin 8)	no	N
biomarkers	il_10_0 (Interleukin 10)	no	N
biomarkers	tyru0 (Tyrosol)	no	N
biomarkers	ohtyru0 (OHTyrosol)	no	N
biomarkers	mohtyu0 (MOH_Tyrosol)	no	N

N:numerical Q: qualitative B:binary

Table 5.10: Attributes included in the dataset of the Initial Characteristics of Volunteers

The habits block ($H = \mathcal{B}^2$) related with the life-style contains the diet habits are described through 14 categorical attributes (see Table 5.11); and physical activity in 10 numerical attributes (see Table 5.12).

	Name	Type	Description
1	p14_1_1	Q	VISIT P14 1: Do you use olive oil as the main cooking fat? INITIAL
2	p14_2_1	Q	VISIT P14 1: How much olive oil consume in total per day? (Including that used for frying, eating out, salads, etc) INITIAL
3	p14_3_1	Q	VISIT P14 1: How many portions of vegetables you eat per day? (side dish or accompaniments count as half portion) 1 portion = 200g INITIAL
4	p14_4_1	Q	VISIT P14 1: How many pieces of fruit (including natural juices) per day? INITIAL

5	p14_5_1	Q	VISIT P14 1: How many portions of red meat, burgers, hot dogs or sausages cosume per day? 1 portion = INITIAL 100-150g
6	p14_6_1	Q	VISIT P14 1: How many portions of butter, margarine or cream per day? INITIAL portion = 12g
7	p14_7_1	Q	VISIT P14 1: How many (sugared) soft drinks (sodas, colas, tonics, bitter) per day? INITIAL
8	p14_8_1	Q	VISIT P14 1: Do you drink wine? How much consumes per week? INITIAL
9	p14_9_1	Q	VISIT P14 1: How many portions of legumes consumes per week? 1 bowl or portion = 150g INITIAL
10	p14_10_1	Q	VISIT P14 1: How many portions of fish/seafood consumes per week? 1 portion = 100-150g of fish or 4-5 pieces or 200gr of seafood INITIAL
11	p14_11_1	Q	VISIT P14 1: How often consume (not homemade) Commercial bakery like cookies, puddings, sweets or cakes per week? INITIAL
12	p14_12_1	Q	VISIT P14 1: How often consume nuts per week? INITIAL 1 portion = 30g
13	p14_13_1	Q	VISIT P14 1: Do you preferably eat meat of chicken, rabbit or turkey instead of beef, pork, burgers and sausages? Chicken: 1 piece or portion 100-150g INITIAL
14	p14_14_1	Q	VISIT P14 1: How many times per week consume cooked vegetables, pasta, rice or other dishes dressed with homemade tomato sauce (garlic, onion or leek made simmered with olive oil)? INITIAL

Table 5.11: Diet Habits Attributes

	Name	Type	Description
1	ge2_3_5s_1	N	AF VISIT 1: light physical activity (last week) KCAL/day
2	ge4_5_5s_1	N	AF VISIT 1: moderate physical activity (last week) KCAL/day
3	ge6_12s_1	N	AF VISIT 1: intense physical activity (last week) KCAL/day
4	gehos_1	N	AF VISIT 1: Physical Activity home (last week) KCAL/day
5	getots_1	N	AF VISIT 1: Total physical activity (last week) KCAL/day
6	ge2_3_5a_1	N	AF VISIT 1: light physical activity (last year) KCAL/day
7	ge4_5_5a_1	N	AF VISIT 1: moderate physical activity (last year) KCAL/day
8	ge6_12a_1	N	AF VISIT 1: Intense physical activity (last year) KCAL/day
9	gehoa_1	N	AF VISIT 1: Physical Activity home (last year) KCAL/day
10	getota_1	N	AF VISIT 1: Total physical activity (last year) KCAL/day

Table 5.12: Physical Activity Attributes

From the 10 physical activity attributes, 4 attributes are included in our analysis because the rest contains redundant information for our purposes. The level of physical activity is asked "per week" and "per year". In addition, *getots_1* (totalWeek) and *getota_1* (totalYear) recollect the sum of the detailed attributes, "per week" and "per year" respectively. Then, the 4 attributes that are referred to the physical activity per week are selected without including the total sum.

The gene expression attributes are not included in those blocks. These attributes are used later according to the proposed methodology in Section 3.2.

5.4 Profiling Initial State of Individuals

According to the proposed methodology in Section 3.2, subjects were clustered twice, using one block per analysis, under the schema of Figure 5.5 and using Ward's method with the Gower's Dissimilarity Coefficient. The hierarchical clustering has been performed for both clusterings in order to know how many clusters exists (see Section 2.3.1). The resulting clusters (P_{C_o}, P_{H_o}) were crossed in order to obtain the final partition (P_o) (see Section 5.4.2). In this case, a simple two way contingency table worked (Cartesian product), as only 2 blocks were involved (see Figure 5.5). Resulting groups gave the profiles of the individuals regarding their baseline characteristics and their habits on diet and physical activity (see Section 5.4.2).

5.4.1 Clustering of the Baseline Characteristics Block (C)

Hierarchical clustering with Ward's method has been applied to the dataset composed by the attributes of the thematic block C and using the Gower's metrics. Figure 5.6 depicts the resulting dendrogram.

Figure 5.7a is the histogram of level indexes of the dendrogram (Figure 5.6). From this picture, it possible to determine a convenient cut in 2 or 3 clusters according to max leaps in the graphic.

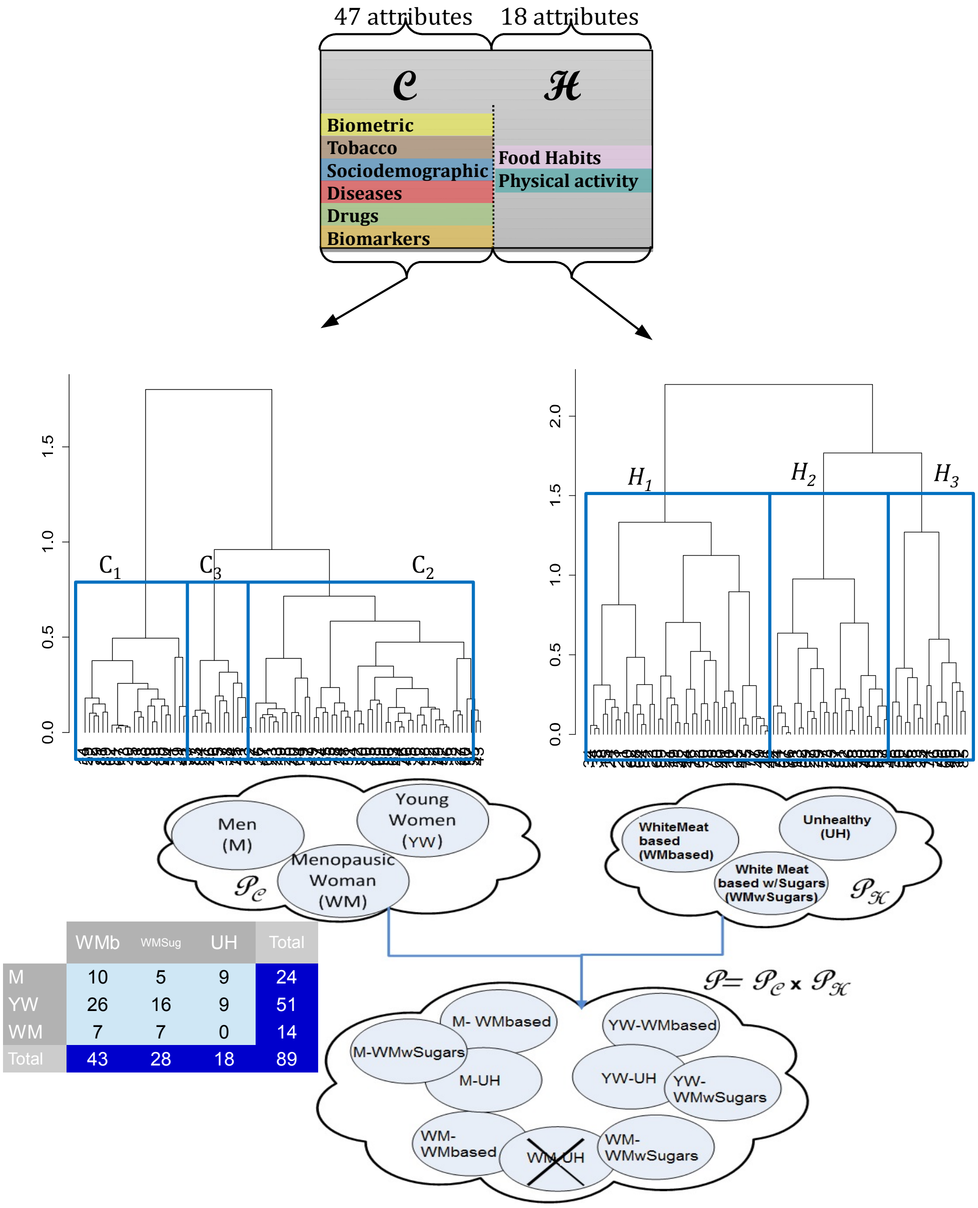
Finally, we selected 3 clusters: $P_{C_o} = \{C_1, C_2, C_3\}$. Figure 5.7b shows the cut in these 3 clusters over the dendrogram.

Table 5.13 shows the basic information of the resulting partition: clusters sizes and the individuals of each cluster.

Table 5.13: Partition P_{C_o}

Cluster	Size	Individuals
C_1	24	$\{i_1, i_{12}, i_{13}, i_{14}, i_{15}, i_{18}, i_{20}, i_{30}, i_{31}, i_{39}, i_{41}, i_{48}, i_{53}, i_{54}, i_{59}, i_{62}, i_{66}, i_{67}, i_{71}, i_{79}, i_{81}, i_{85}, i_{88}, i_{89}\}$
C_2	51	$\{i_2, i_3, i_4, i_5, i_6, i_7, i_8, i_9, i_{10}, i_{19}, i_{22}, i_{23}, i_{24}, i_{27}, i_{28}, i_{29}, i_{32}, i_{33}, i_{34}, i_{35}, i_{36}, i_{37}, i_{40}, i_{42}, i_{43}, i_{44}, i_{45}, i_{46}, i_{47}, i_{49}, i_{50}, i_{51}, i_{52}, i_{55}, i_{56}, i_{57}, i_{58}, i_{60}, i_{61}, i_{63}, i_{64}, i_{65}, i_{68}, i_{69}, i_{70}, i_{72}, i_{76}, i_{78}, i_{80}, i_{86}, i_{87}\}$
C_3	14	$\{i_{11}, i_{16}, i_{17}, i_{21}, i_{25}, i_{26}, i_{38}, i_{73}, i_{74}, i_{75}, i_{77}, i_{82}, i_{83}, i_{84}\}$

After the characterization of the partition in Section 5.4.3, the classes are renamed as: Men (M), Young Women (YW) and Women with Menopause (WM).



	WMb	WMSug	UH	Total
M	10	5	9	24
YW	26	16	9	51
WM	7	7	0	14
Total	43	28	18	89

Figure 5.5: Integrative Multiview Clustering Scheme applied to the case study before the intervention

.

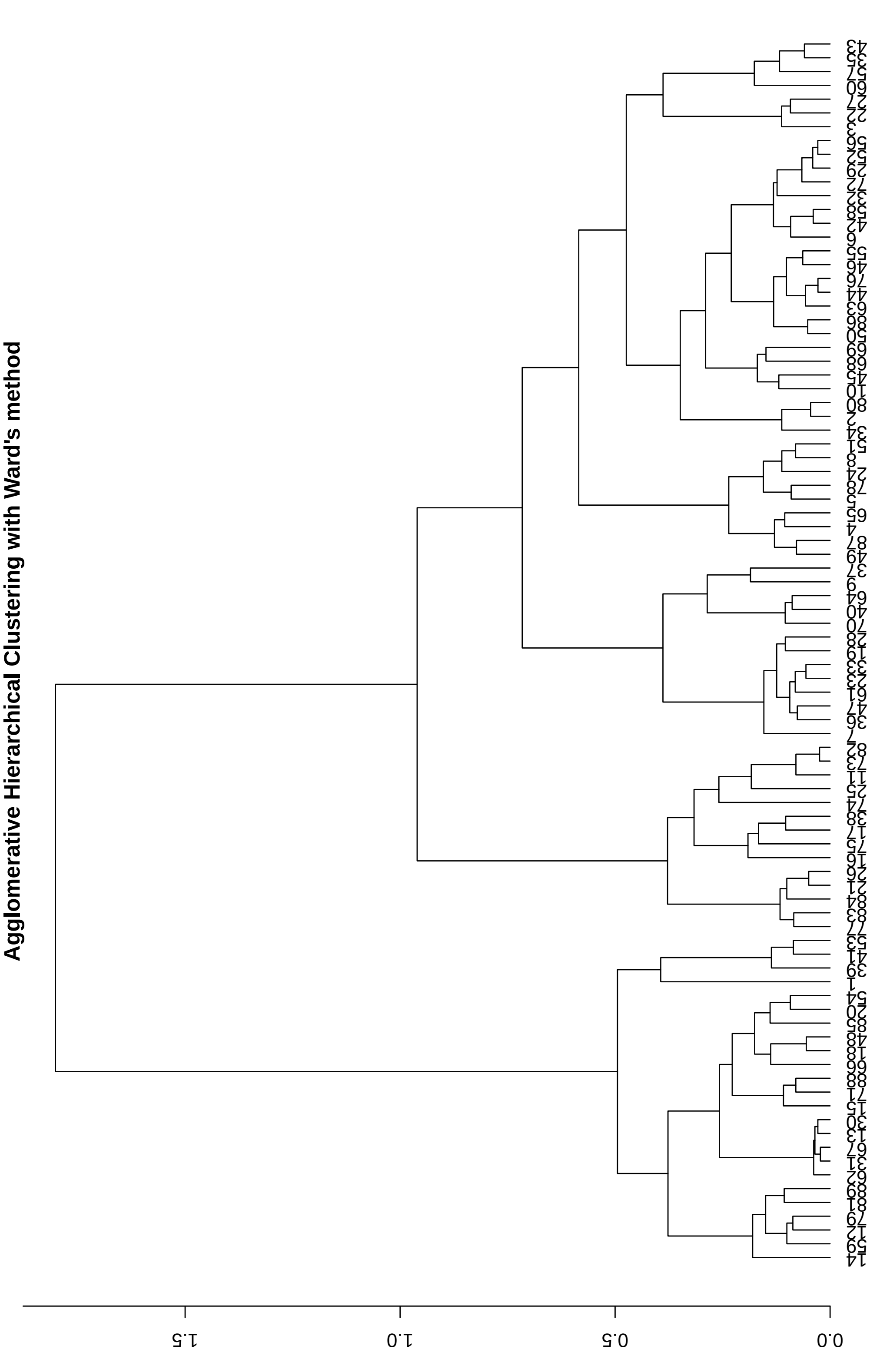
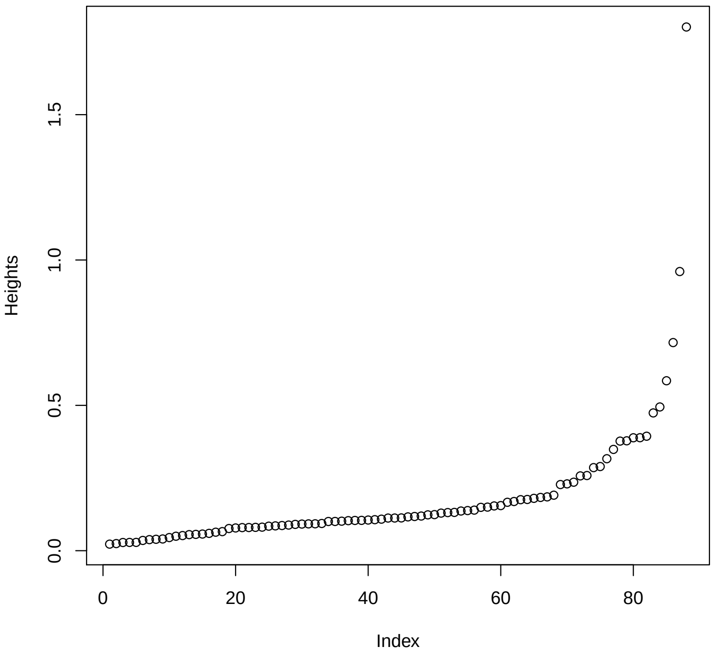
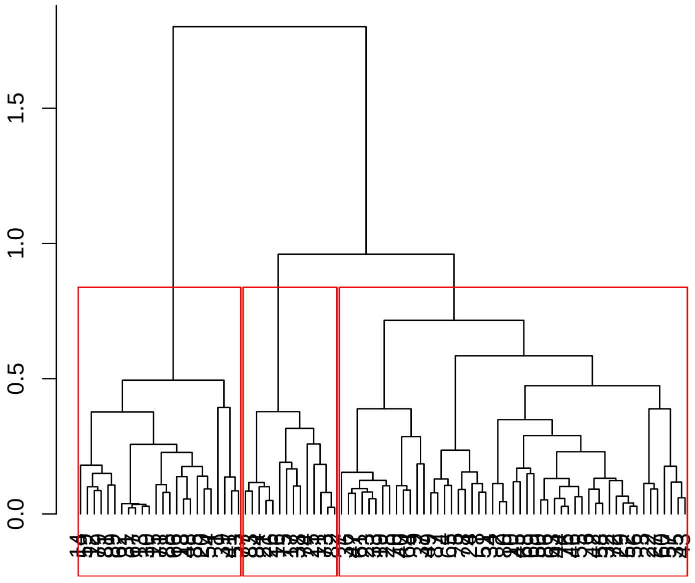


Figure 5.6: Dendrogrem of the Hierarchical clustering with Ward's Method for baseline characteristics (C)



(a) Heights of level indexes



(b) Dendrogram cut in 3 clusters

Figure 5.7: Hierarchical clustering with Ward's Method for baseline characteristics

5.4.1.1 Clustering of Diet and Physical Activity Habits Block (H)

The second clustering is applied over the attributes referred to the diet and physical activity at the beginning of the study. These attributes belong to the diet (p14-item) and physical activity questionnaire of the first visit in the study (see Section 5.3).

The dendrogram of the resulting hierarchical clustering with Ward's method and using the Gower's metrics is depicted in Figure 5.8. In Figure 5.9a, the histogram of level indexes of the dendrogram is shown. From this picture, it is possible to determine a convenient cut in 3 clusters according to max leaps in the graphic: $P_{H_o} = \{H_1, H_2, H_3\}$.

Figure 5.9b shows the cut in three clusters over the dendrogram of the resulting hierarchical clustering.

In Table 5.14 is shown how the individuals are split into the 3 classes.

Table 5.14: Resulting partition P_{H_o}

Cluster	Size	Individuals
H_1	43	i_1, i_2, i_3, i_4, i_5, i_7, i_8, i_{10}, i_{11}, i_{15}, i_{16}, i_{19}, i_{20}, i_{21}, i_{22}, i_{23}, i_{24}, i_{30}, i_{31}, i_{34}, i_{37}, i_{38}, i_{40}, i_{41}, i_{42}, i_{46}, i_{47}, i_{49}, i_{54}, i_{55}, i_{60}, i_{62}, i_{63}, i_{64}, i_{65}, i_{68}, i_{72}, i_{75}, i_{77}, i_{78}, i_{84}, i_{88}, i_{89}
H_2	28	i_9, i_{13}, i_{17}, i_{25}, i_{26}, i_{27}, i_{28}, i_{29}, i_{39}, i_{43}, i_{44}, i_{51}, i_{52}, i_{53}, i_{57}, i_{58}, i_{61}, i_{67}, i_{69}, i_{70}, i_{73}, i_{74}, i_{76}, i_{80}, i_{81}, i_{82}, i_{83}, i_{87}
H_3	18	i_6, i_{12}, i_{14}, i_{18}, i_{32}, i_{33}, i_{35}, i_{36}, i_{45}, i_{48}, i_{50}, i_{56}, i_{59}, i_{66}, i_{71}, i_{79}, i_{85}, i_{86}

After the characterization of the partition in Section 5.4.4, the classes are renamed as: Based on White Meat (WMbased), White Meat based with Sugars (WMwSugars) and Unhealthy (UH).

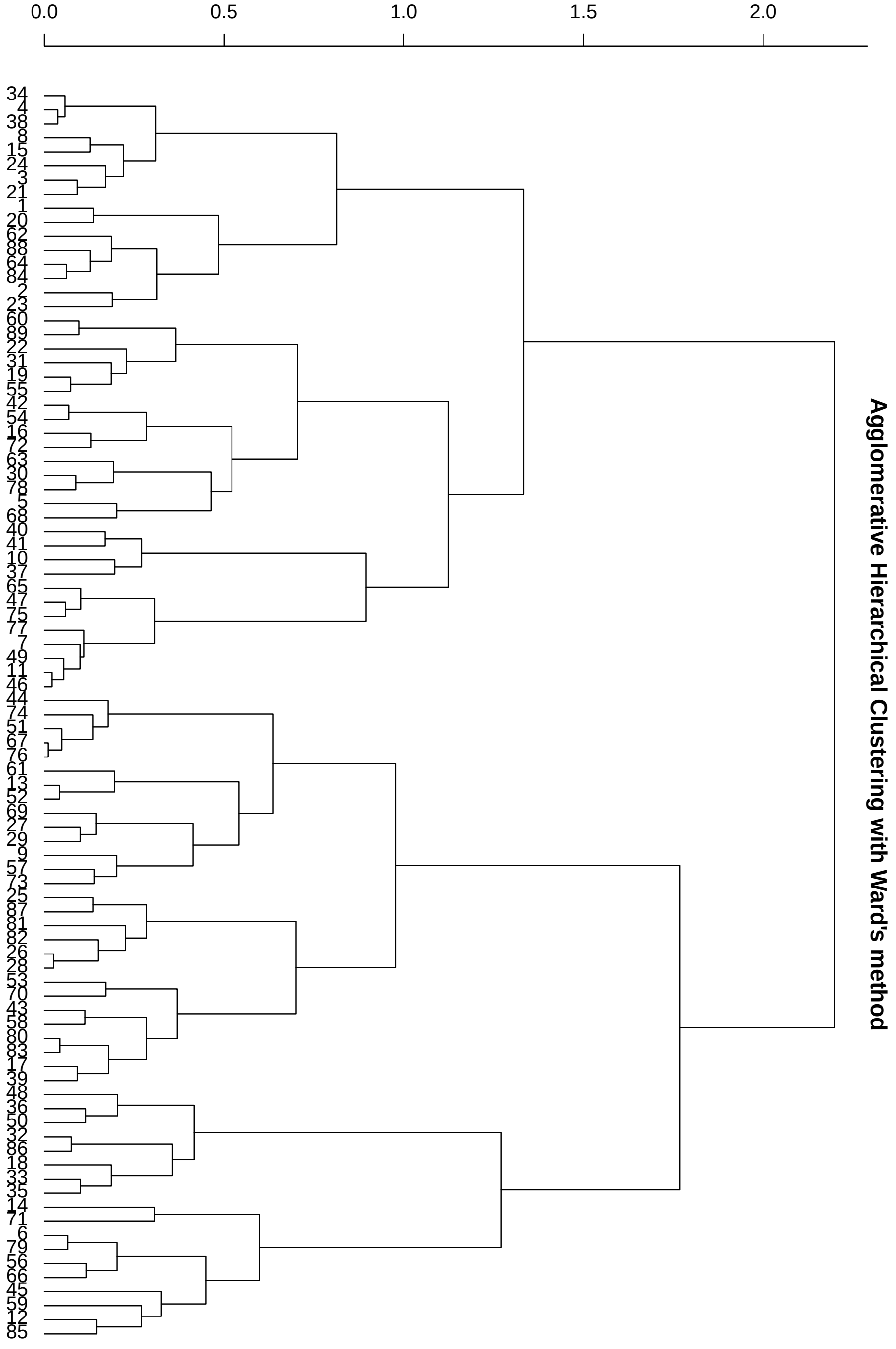
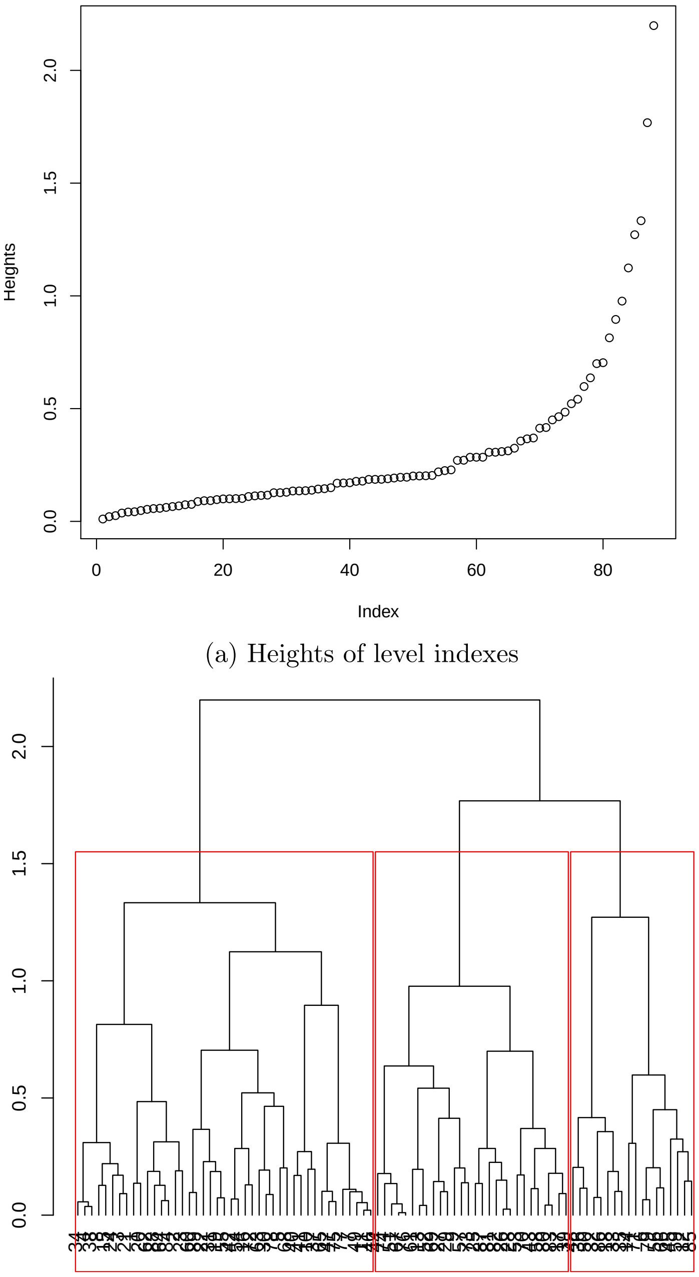


Figure 5.8: Hierarchical clustering with Ward's Method for Diet and Physical Activity Habits



(a) Heights of level indexes

(b) Dendrogram cut in 3 clusters

Figure 5.9: Hierarchical clustering with Ward's Method for habits (Diet and Physical Activity) at the beginning of the study

5.4.2 Building the Cross Clustering with Baseline Characteristics and Habits

In next Table 5.15, the clusters of both baseline characteristics ($P_{C_o} = \{M, YW, WM\}$) and habits ($P_{H_o} = \{WMbased, WMwSugars, UH\}$) clusterings are crossed ($P_o = P_{C_o} \times P_{H_o}$).

Table 5.15: Crossing clusters

	WMbased	WMwSugars	UH	Total
M	10	5	9	24
YW	26	16	9	51
WM	7	7	0	14
Total	28	43	18	89

From this Table 5.15, new clusters (P_o) can be subtracted as the Cartesian product of P_{C_o} and P_{H_o}. The Table 5.15 shows the clusters size. The class $WM \times UH$ is null which means that no subjects in WM has the diet and physical activity habits described by UH for current sample. Therefore, there are 8 cross clusters to be considered. The names of these 8 clusters are the following:

1. M-WMbased
2. M-WMwSugars
3. M-UH
4. WY-WMbased
5. WY-WMwSugars
6. WY-UH
7. WM-WMbased
8. WM-WMwSugars

5.4.3 Characterization and Interpretation of the partition P_{C_o}

In this section, the attributes used for the clustering (C) are analyzed to characterize the 3 resulting clusters of P_{C_o}. The Cluster Interpretation Methodology CI-IMS proposed in Section 3.6 has been applied over the resulting partition P_{C_o}.

Given $\mathcal{C}$ the set of attributes of the thematic block C and partition P_{C_o},

1. Set $\delta = 0.25 \in [0, 1]$ admitting till 25% of contamination to consider a class described by a value.
2. Set $\alpha = 0.01 \in [0, 1]$ among habitual values in hypothesis tests.
3. Set $\varepsilon_1 = 0.3,\ \varepsilon_2 = 0.2 \in (0, 0.5] \wedge \varepsilon_1 + \varepsilon_2 < 1$ to make the Sensitive Analysis with $\{0.5n, 0.7n, n, 1.3n, 1.5n\}$.
4. Set the minimum level of robustness $r = "W" \in dp$.
5. Build $\mathcal{V}_P$ by applying $\forall X \in \mathcal{C} SA - MTV(X, P, \varepsilon_1, \varepsilon_2, \alpha, \delta)$ defined in Section 3.5. Tables 5.16, 5.17 and 5.18 show the results of applying SA-MTV to each attribute $X \in \mathcal{C}$ over each class.

6. The interpretation of P_{C_o} at level $r =$ "W" is $\Upsilon_{P_{C_o}, "W"}$: the automatic generated profiles for class M is the following (the rest of profiles are in Appendix B.1.2 in verbose format):

   Profile Class M:

   Biometrics attributes:

   - **sexo (gender) - Woman: Less (B)**
   - altura1 (height): Higher (R)
   - peso1 (weight): Higher (R)
   - imc1 (BMI): Higher (R)
   - cint1 (waist): Higher (R)
   - pas_esds_2 (systolic Pressure): Higher (R)
   - pad_esds_2 (diastolic Pressure): Higher (M)
   - fc_d_2 (heart Rate): Lower (R)

   Menopause & Diseases attributes:

   - **menopausia_bin (menopause) - TRUE: Less (B)**
   - pertensa (stressful): Lower (R)
   - depre (depression) - TRUE: Less (B)
   - cancer (cancer) - TRUE: Less (B)
   - fractura (bone fracture) - TRUE: Less (B)
   - disnea (dyspnea) - TRUE: Less (B)

   Drugs attributes:

   - maspirina1 (aspirin) - TRUE: Less (B)
   - **mdolor1 (NSAID) - TRUE: Less (B)**
   - **mansiedad1 (anxiolytic) - TRUE: Less (B)**
   - **mvitMin1 (vitamin or minerals) - TRUE: Less (B)**
   - mtension1 (heart) - TRUE: Less (B)
   - **mhormo1 (hormon) - TRUE: Less (B)**
   - motros1 (other) - TRUE: Less (B)

   Biomarkers attributes:

   - gluc0 (glucose): Higher (W)
   - cholest0 (cholesterol): Lower (W)
   - hdl0 (HDL): Lower (R)
   - tryg0 (triglycerides): Higher (R)
   - isoprostanos_0 (F2 α Isoprostanes): Higher (M)

- mcp_1_0 (Monocyte Chemotactic Protein-1): Higher (R)
- t_pa_0 (Tissue Plasminogen Activator): Higher (R)

7. Visualize the Class Panel Graphs using only the subset of attributes that appear in $\Upsilon_{P_{C_o},"W"}$. 38 attributes are descriptors for at least one class, the corresponding CPG are shown in Table 5.19, Table 5.20, Table 5.21, Table 5.22, Table 5.23, Table 5.24 and Table 5.25.

Table 5.16: Sensitive Analysis of Class *Man* (M)

		Sample size						
		ν_1	ν_2	ν_3	ν_4	ν_5		
		$n(1-\varepsilon_1-\varepsilon_2)$	$n(1-\varepsilon_1)$	n	$n(1+\varepsilon_1)$	$n(1+\varepsilon_1+\varepsilon_2)$		
Attribute	**Category**	44	62	89	115	133	**dp**	**sense**
sexo (gender)	Woman	B	B	B	B	B	B	⊥
EDAD (Age)		7.80e-02	4.08e-02	1.54e-02	6.01e-03*	3.14e-03*	$\overline{W}$	↓
altura1 (height)		5.52e-11*	5.87e-15*	6.42e-21*	1.17e-26*	1.24e-30*	R	↑
peso1 (weight)		4.87e-10*	1.25e-13*	5.16e-19*	3.36e-24*	8.63e-28*	R	↑
imc1 (BMI)		1.87e-03*	2.20e-04*	8.80e-06*	3.97e-07*	4.64e-08*	R	↑
cint1 (waist)		2.26e-08*	2.75e-11*	1.16e-15*	7.15e-20*	8.68e-23*	R	↑
pas_esds_2 (systolic Pressure)		7.79e-06*	1.00e-07*	1.47e-10*	2.73e-13*	3.51e-15*	R	↑
pad_esds_2 (diastolic Pressure)		2.25e-02	7.17e-03*	1.28e-03*	2.45e-04*	7.76e-05*	M	↑
fc_d_2 (heart Rate)		8.41e-03*	1.80e-03*	1.78e-04*	1.92e-05*	4.10e-06*	R	↓
tabaco (tobacco)	ex>5	B	B	B	B	B	B	⊥
tabaco (tobacco)	ex0-1	B	B	B	B	B	B	⊥
tabaco (tobacco)	ex1-5	B	B	B	B	B	B	⊥
tabaco (tobacco)	Never	3.29e-01	3.04e-01	2.70e-01	2.40e-01	2.22e-01	$\overline{R}$	↓
tabaco (tobacco)	Yes	B	B	B	B	B	B	⊥
est_civi (civilState)	casado/a	3.91e-01	3.88e-01	3.83e-01	3.79e-01	3.76e-01	$\overline{R}$	↑
est_civi (civilState)	Divorciado/a	B	B	B	B	B	B	⊥
est_civi (civilState)	Separado/a	B	B	B	B	B	B	⊥
est_civi (civilState)	soltero	3.57e-01	3.41e-01	3.19e-01	2.99e-01	2.86e-01	$\overline{R}$	↑
est_civi (civilState)	Viudo/a	B	B	B	B	B	B	⊥
menopausia_bin (menopause)	TRUE	B	B	B	B	B	B	⊥
pertensa (stressful)		2.34e-06*	1.86e-08*	1.32e-11*	1.21e-14*	9.62e-17*	R	↓
depre (depression)	TRUE	B	B	B	B	B	B	⊥
cancer (cancer)	TRUE	B	B	B	B	B	B	⊥
fractura (bone fracture)	TRUE	B	B	B	B	B	B	⊥
disnea (dyspnea)	TRUE	B	B	B	B	B	B	⊥
maspirina1 (aspirin)	TRUE	B	B	B	B	B	B	⊥
mdolor1 (NSAID)	TRUE	B	B	B	B	B	B	⊥
mansiedad1 (anxiolytic)	TRUE	B	B	B	B	B	B	⊥
mvitMin1 (vitamin or minerals)	TRUE	B	B	B	B	B	B	⊥
mtension1 (heart)	TRUE	B	B	B	B	B	B	⊥
mhormo1 (hormon)	TRUE	B	B	B	B	B	B	⊥
motros1 (other)	TRUE	B	B	B	B	B	B	⊥
gluc0 (glucose)		6.42e-02	3.10e-02	1.04e-02*	3.64e-03*	1.76e-03*	W	↑

Attribute	**Category**	$\nu_1 = 44$	$\nu_2 = 62$	$\nu_3 = 89$	$\nu_4 = 115$	$\nu_5 = 133$	**dp**	**sense**
cholest0 (cholesterol)		6.70e-02	3.30e-02	1.14e-02*	4.07e-03*	2.00e-03*	W	↓
LDL_0 (LDL)		1.00e-01	5.77e-02	2.53e-02	1.14e-02*	6.58e-03*	$\overline{W}$	↓
hdl0 (HDL)		4.33e-05*	1.12e-06*	4.60e-09*	2.32e-11*	5.97e-13*	R	↓
tryg0 (triglycerides)		5.19e-04*	3.63e-05*	6.71e-07*	1.44e-08*	1.00e-09*	R	↑
oxldl0 (Oxidized LDL)		3.61e-01	3.49e-01	3.30e-01	3.13e-01	3.02e-01	$\overline{R}$	↓
isoprostanos_0 (F2 α Isoprostanes)		1.74e-02	4.99e-03*	7.64e-04*	1.25e-04*	3.59e-05*	M	↑
ifn_g_0 (Interferon-γ)		3.23e-01	2.99e-01	2.65e-01	2.35e-01	2.17e-01	$\overline{R}$	↓
mcp_1_0 (Monocyte Chemotactic Protein-1)		1.81e-03*	2.09e-04*	8.20e-06*	3.62e-07*	4.18e-08*	R	↑
s_p_selectin_0 (sP-selectin)		1.02e-01	5.97e-02	2.65e-02	1.21e-02*	7.06e-03*	$\overline{W}$	↑
s_cd40l_0 (sCD40 Ligand)		3.72e-01	3.64e-01	3.50e-01	3.38e-01	3.30e-01	$\overline{R}$	↑
pcr0 (C-Reactive Protein)		2.41e-01	1.98e-01	1.47e-01	1.11e-01	9.08e-02	$\overline{R}$	↓
oxo_gg_0 (8-Oxoguanine)		3.88e-01	3.86e-01	3.81e-01	3.77e-01	3.74e-01	$\overline{R}$	↓
tnf_a_0 (Tumor necrosis factor-α)		2.20e-01	1.74e-01	1.22e-01	8.68e-02	6.86e-02	$\overline{R}$	↓
t_pa_0 (Tissue Plasminogen Activator)		5.62e-04*	4.05e-05*	7.85e-07*	1.76e-08*	1.27e-09*	R	↑
s_vcam_1_0 (Soluble cell adhesion molecules-1)		8.63e-02	4.70e-02	1.88e-02	7.81e-03*	4.25e-03*	$\overline{W}$	↓
il_6_0 (Interleukin 6)		8.09e-02	4.29e-02	1.66e-02	6.62e-03*	3.50e-03*	$\overline{W}$	↓
il_8_0 (Interleukin 8)		1.91e-01	1.43e-01	9.25e-02	6.07e-02	4.53e-02	$\overline{R}$	↓
il_10_0 (Interleukin 10)		2.06e-01	1.59e-01	1.07e-01	7.34e-02	5.64e-02	$\overline{R}$	↓
tyru0 (Tyrosol)		2.26e-01	1.80e-01	1.29e-01	9.29e-02	7.42e-02	$\overline{R}$	↑
ohtyru0 (OHTyrosol)		1.79e-01	1.31e-01	8.15e-02	5.16e-02	3.76e-02	$\overline{R}$	↑
mohtyu0 (MOH_Tyrosol)		2.66e-01	2.28e-01	1.79e-01	1.43e-01	1.22e-01	$\overline{R}$	↑
✓: $p - value < \alpha$		test $\in \Theta$						

Table 5.17: Sensitive Analysis of Class *Young Woman* (YW)

		Sample size						
		ν_1	ν_2	ν_3	ν_4	ν_5		
		$n(1-\varepsilon_1-\varepsilon_2)$	$n(1-\varepsilon_1)$	n	$n(1+\varepsilon_1)$	$n(1+\varepsilon_1+\varepsilon_2)$		
Attribute	**Category**	44	62	89	115	133	**dp**	**sense**
sexo (gender)	Woman	B	B	B	B	B	B	⊤
EDAD (Age)		2.27e-02	7.24e-03*	1.30e-03*	2.49e-04*	7.94e-05*	M	↓
altura1 (height)		2.01e-03*	2.42e-04*	1.01e-05*	4.74e-07*	5.71e-08*	R	↓
peso1 (weight)		1.88e-04*	8.73e-06*	8.73e-08*	1.03e-09*	4.79e-11*	R	↓
imc1 (BMI)		1.43e-02	3.78e-03*	5.14e-04*	7.52e-05*	1.99e-05*	M	↓
cint1 (waist)		8.20e-05*	2.73e-06*	1.65e-08*	1.21e-10*	4.02e-12*	R	↓
pas_esds_2 (systolic Pressure)		1.81e-04*	8.28e-06*	8.10e-08*	9.39e-10*	4.29e-11*	R	↓
pad_esds_2 (diastolic Pressure)		2.26e-03*	2.85e-04*	1.28e-05*	6.43e-07*	8.10e-08*	R	↓
fc_d_2 (heart Rate)		3.73e-01	3.64e-01	3.52e-01	3.39e-01	3.31e-01	$\overline{R}$	↑
tabaco (tobacco)	ex>5	B	B	B	B	B	B	⊥

Attribute	**Category**	$\nu_1 = 44$	$\nu_2 = 62$	$\nu_3 = 89$	$\nu_4 = 115$	$\nu_5 = 133$	**dp**	**sense**
tabaco (tobacco)	ex0-1	B	B	B	B	B	B	$\perp$
tabaco (tobacco)	ex1-5	B	B	B	B	B	B	$\perp$
tabaco (tobacco)	Never	3.98e-01	3.97e-01	3.96e-01	3.96e-01	3.95e-01	$\overline{R}$	$\uparrow$
tabaco (tobacco)	Yes	2.36e-01	1.90e-01	1.38e-01	1.01e-01	8.12e-02	$\overline{R}$	$\uparrow$
est_civi (civilState)	casado/a	3.21e-01	2.93e-01	2.56e-01	2.25e-01	2.06e-01	$\overline{R}$	$\downarrow$
est_civi (civilState)	Divorciado/a	B	B	B	B	B	B	$\perp$
est_civi (civilState)	Separado/a	B	B	B	B	B	B	$\perp$
est_civi (civilState)	soltero	1.92e-01	1.42e-01	9.05e-02	5.87e-02	4.35e-02	$\overline{R}$	$\uparrow$
est_civi (civilState)	Viudo/a	B	B	B	B	B	B	$\perp$
menopausia_bin (menopause)	TRUE	B	B	B	B	B	B	$\perp$
pertensa (stressful)		6.74e-03*	1.32e-03*	1.14e-04*	1.08e-05*	2.11e-06*	R	$\uparrow$
depre (depression)	TRUE	B	B	B	B	B	B	$\perp$
cancer (cancer)	TRUE	B	B	B	B	B	B	$\perp$
fractura (bone fracture)	TRUE	B	B	B	B	B	B	$\perp$
disnea (dyspnea)	TRUE	B	B	B	B	B	B	$\perp$
maspirina1 (aspirin)	TRUE	B	B	B	B	B	B	$\perp$
mdolor1 (NSAID)	TRUE	B	B	B	B	B	B	$\perp$
mansiedad1 (anxiolytic)	TRUE	B	B	B	B	B	B	$\perp$
mvitMin1 (vitamin or minerals)	TRUE	B	B	B	B	B	B	$\perp$
mtension1 (heart)	TRUE	B	B	B	B	B	B	$\perp$
mhormo1 (hormon)	TRUE	B	B	B	B	B	B	$\perp$
motros1 (other)	TRUE	3.05e-01	2.73e-01	2.31e-01	1.97e-01	1.77e-01	$\overline{R}$	$\uparrow$
gluc0 (glucose)		6.39e-03*	1.23e-03*	1.03e-04*	9.44e-06*	1.81e-06*	R	$\downarrow$
cholest0 (cholesterol)		2.62e-01	2.23e-01	1.74e-01	1.37e-01	1.16e-01	$\overline{R}$	$\downarrow$
LDL_0 (LDL)		2.12e-01	1.65e-01	1.14e-01	7.91e-02	6.16e-02	$\overline{R}$	$\downarrow$
hdl0 (HDL)		1.51e-02	4.08e-03*	5.73e-04*	8.64e-05*	2.33e-05*	M	$\uparrow$
tryg0 (triglycerides)		1.91e-03*	2.26e-04*	9.14e-06*	4.17e-07*	4.91e-08*	R	$\downarrow$
oxldl0 (Oxidized LDL)		5.82e-02	2.71e-02	8.56e-03*	2.83e-03*	1.31e-03*	W	$\downarrow$
isoprostanos_0 (F2 α Isoprostanes)		3.66e-01	3.55e-01	3.39e-01	3.23e-01	3.13e-01	$\overline{R}$	$\downarrow$
ifn_g_0 (Interferon-γ)		3.08e-01	2.78e-01	2.39e-01	2.07e-01	1.87e-01	$\overline{R}$	$\uparrow$
mcp_1_0 (Monocyte Chemotactic Protein-1)		1.30e-01	8.30e-02	4.25e-02	2.23e-02	1.42e-02	$\overline{R}$	$\downarrow$
s_p_selectin_0 (sP-selectin)		3.91e-01	3.90e-01	3.87e-01	3.84e-01	3.82e-01	$\overline{R}$	$\downarrow$
s_cd40l_0 (sCD40 Ligand)		2.83e-01	2.47e-01	2.02e-01	1.66e-01	1.45e-01	$\overline{R}$	$\uparrow$
pcr0 (C-Reactive Protein)		3.73e-01	3.65e-01	3.52e-01	3.40e-01	3.32e-01	$\overline{R}$	$\uparrow$
oxo_gg_0 (8-Oxoguanine)		3.90e-01	3.89e-01	3.85e-01	3.82e-01	3.80e-01	$\overline{R}$	$\uparrow$
tnf_a_0 (Tumor necrosis factor-α)		2.13e-01	1.67e-01	1.15e-01	8.04e-02	6.27e-02	$\overline{R}$	$\uparrow$
t_pa_0 (Tissue Plasminogen Activator)		6.68e-03*	1.30e-03*	1.12e-04*	1.06e-05*	2.06e-06*	R	$\downarrow$
s_vcam_1_0 (Soluble cell adhesion molecules-1)		1.01e-01	5.82e-02	2.56e-02	1.16e-02*	6.70e-03*	$\overline{W}$	$\uparrow$
il_6_0 (Interleukin 6)		2.15e-01	1.69e-01	1.17e-01	8.24e-02	6.45e-02	$\overline{R}$	$\uparrow$
il_8_0 (Interleukin 8)		3.88e-01	3.85e-01	3.80e-01	3.75e-01	3.72e-01	$\overline{R}$	$\downarrow$
il_10_0 (Interleukin 10)		2.99e-01	2.68e-01	2.26e-01	1.92e-01	1.72e-01	$\overline{R}$	$\downarrow$

Attribute	**Category**	$\nu_1 = 44$	$\nu_2 = 62$	$\nu_3 = 89$	$\nu_4 = 115$	$\nu_5 = 133$	**dp**	**sense**
tyru0 (Tyrosol)		3.09e-01	2.81e-01	2.42e-01	2.10e-01	1.90e-01	$\overline{R}$	↑
ohtyru0 (OHTyrosol)		3.94e-01	3.93e-01	3.92e-01	3.91e-01	3.90e-01	$\overline{R}$	↓
mohtyu0 (MOH_Tyrosol)		3.94e-01	3.94e-01	3.93e-01	3.93e-01	3.92e-01	$\overline{R}$	↑
✓: $p - value < \alpha$		test $\in \Theta$						

Table 5.18: Sensitive Analysis of Class *Woman with Menopause* (WM)

		Sample size						
		ν_1	ν_2	ν_3	ν_4	ν_5		
		$n(1-\varepsilon_1-\varepsilon_2)$	$n(1-\varepsilon_1)$	n	$n(1+\varepsilon_1)$	$n(1+\varepsilon_1+\varepsilon_2)$		
Attribute	**Category**	44	62	89	115	133	**dp**	**sense**
sexo (gender)	Woman	B	B	B	B	B	B	⊤
EDAD (Age)		1.43e-11*	8.80e-16*	4.23e-22*	3.49e-28*	2.14e-32*	R	↑
altura1 (height)		8.61e-06*	1.16e-07*	1.80e-10*	3.54e-13*	4.74e-15*	R	↓
peso1 (weight)		5.38e-02	2.42e-02	7.31e-03*	2.30e-03*	1.04e-03*	W	↓
imc1 (BMI)		1.54e-01	1.06e-01	6.02e-02	3.49e-02	2.39e-02	$\overline{R}$	↑
cint1 (waist)		3.46e-01	3.28e-01	3.03e-01	2.80e-01	2.65e-01	$\overline{R}$	↑
pas_esds_2 (systolic Pressure)		1.42e-02	3.75e-03*	5.08e-04*	7.41e-05*	1.95e-05*	M	↑
pad_esds_2 (diastolic Pressure)		2.73e-05*	5.84e-07*	1.82e-09*	7.03e-12*	1.50e-13*	R	↑
fc_d_2 (heart Rate)		9.22e-04*	8.13e-05*	2.12e-06*	6.34e-08*	5.58e-09*	R	↑
tabaco (tobacco)	ex>5	B	B	B	B	B	B	⊥
tabaco (tobacco)	ex0-1	B	B	B	B	B	B	⊥
tabaco (tobacco)	ex1-5	B	B	B	B	B	B	⊥
tabaco (tobacco)	Never	2.93e-01	2.58e-01	2.14e-01	1.78e-01	1.57e-01	$\overline{R}$	↑
tabaco (tobacco)	Yes	B	B	B	B	B	B	⊥
est_civi (civilState)	casado/a	B	B	B	B	B	B	⊤
est_civi (civilState)	Divorciado/a	8.00e-05*	2.46e-06*	1.32e-08*	8.64e-11*	2.66e-12*	R	↑
est_civi (civilState)	Separado/a	2.60e-01	2.19e-01	1.68e-01	1.31e-01	1.10e-01	$\overline{R}$	↑
est_civi (civilState)	soltero	2.06e-04*	9.33e-06*	8.98e-08*	1.03e-09*	4.65e-11*	R	↓
est_civi (civilState)	Viudo/a	2.31e-02	7.21e-03*	1.26e-03*	2.33e-04*	7.27e-05*	M	↑
menopausia_bin (menopause)	TRUE	B	B	B	B	B	B	⊤
pertensa (stressful)		2.82e-01	2.47e-01	2.02e-01	1.66e-01	1.45e-01	$\overline{R}$	↑
depre (depression)	TRUE	B	B	B	B	B	B	⊥
cancer (cancer)	TRUE	2.16e-12*	5.33e-17*	6.51e-24*	1.43e-30*	3.53e-35*	R	↑
fractura (bone fracture)	TRUE	B	B	B	B	B	B	⊥
disnea (dyspnea)	TRUE	B	B	B	B	B	B	⊥
maspirina1 (aspirin)	TRUE	B	B	B	B	B	B	⊥
mdolor1 (NSAID)	TRUE	3.77e-03*	5.59e-04*	3.20e-05*	2.03e-06*	3.02e-07*	R	↑
mansiedad1 (anxiolytic)	TRUE	B	B	B	B	B	B	⊥
mvitMin1 (vitamin or minerals)	TRUE	B	B	B	B	B	B	⊥
mtension1 (heart)	TRUE	B	B	B	B	B	B	⊥
mhormo1 (hormon)	TRUE	B	B	B	B	B	B	⊥
motros1 (other)	TRUE	B	B	B	B	B	B	⊥
gluc0 (glucose)		8.18e-05*	2.72e-06*	1.65e-08*	1.20e-10*	4.00e-12*	R	↑
cholest0 (cholesterol)		5.61e-06*	6.34e-08*	7.59e-11*	1.17e-13*	1.31e-15*	R	↑
LDL_0 (LDL)		3.29e-06*	2.99e-08*	2.60e-11*	2.92e-14*	2.65e-16*	R	↑
hdl0 (HDL)		2.80e-01	2.44e-01	1.98e-01	1.62e-01	1.41e-01	$\overline{R}$	↑

Attribute	Category	$\nu_1 = 44$	$\nu_2 = 62$	$\nu_3 = 89$	$\nu_4 = 115$	$\nu_5 = 133$	dp	sense
tryg0 (triglycerides)		1.91e-02	5.66e-03*	9.16e-04*	1.58e-04*	4.70e-05*	M	↑
oxldl0 (Oxidized LDL)		1.58e-06*	1.07e-08*	5.97e-12*	4.38e-15*	2.96e-17*	R	↑
isoprostanos_0 (F2 α Isoprostanes)		5.21e-03*	9.20e-04*	6.82e-05*	5.56e-06*	9.80e-07*	R	↓
ifn_g_0 (Interferon-γ)		2.59e-01	2.18e-01	1.69e-01	1.32e-01	1.11e-01	$\overline{R}$	↓
mcp_1_0 (Monocyte Chemotactic Protein-1)		4.03e-02	1.62e-02	4.10e-03*	1.09e-03*	4.38e-04*	W	↓
s_p_selectin_0 (sP-selectin)		1.23e-02*	3.07e-03*	3.81e-04*	5.11e-05*	1.27e-05*	R	↓
s_cd40l_0 (sCD40 Ligand)		3.91e-03*	6.16e-04*	3.84e-05*	2.65e-06*	4.17e-07*	R	↓
pcr0 (C-Reactive Protein)		3.08e-01	2.79e-01	2.41e-01	2.08e-01	1.88e-01	$\overline{R}$	↑
oxo_gg_0 (8-Oxoguanine)		3.93e-01	3.92e-01	3.91e-01	3.89e-01	3.88e-01	$\overline{R}$	↓
tnf_a_0 (Tumor necrosis factor-α)		1.77e-01	1.28e-01	7.91e-02	4.97e-02	3.60e-02	$\overline{R}$	↓
t_pa_0 (Tissue Plasminogen Activator)		1.32e-01	8.54e-02	4.42e-02	2.35e-02	1.51e-02	$\overline{R}$	↑
s_vcam_1_0 (Soluble cell adhesion molecules-1)		1.05e-01	6.14e-02	2.76e-02	1.28e-02*	7.50e-03*	$\overline{W}$	↓
il_6_0 (Interleukin 6)		3.92e-01	3.91e-01	3.89e-01	3.87e-01	3.85e-01	$\overline{R}$	↑
il_8_0 (Interleukin 8)		6.30e-03*	1.20e-03*	1.00e-04*	9.11e-06*	1.73e-06*	R	↑
il_10_0 (Interleukin 10)		8.46e-05*	2.85e-06*	1.76e-08*	1.31e-10*	4.42e-12*	R	↑
tyru0 (Tyrosol)		2.15e-03*	2.66e-04*	1.15e-05*	5.63e-07*	6.95e-08*	R	↓
ohtyru0 (OHTyrosol)		1.07e-01	6.36e-02	2.90e-02	1.36e-02*	8.08e-03*	$\overline{W}$	↓
mohtyu0 (MOH_Tyrosol)		1.08e-01	6.44e-02	2.95e-02	1.39e-02	8.29e-03*	$\overline{M}$	↓
✓: $p - value < \alpha$		test $\in \Theta$						

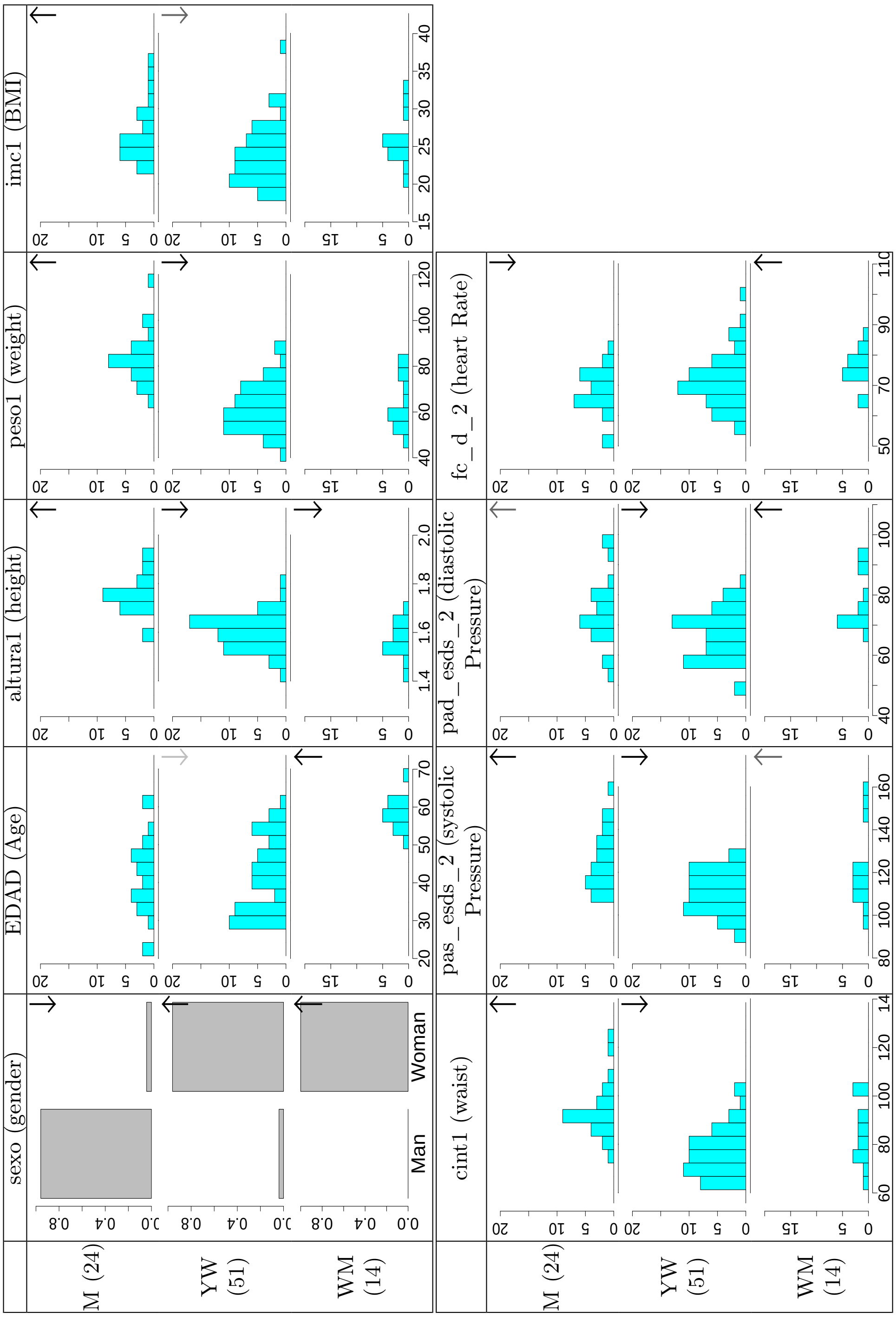


Table 5.19: Biometrics characteristics ~ Baseline (Pre

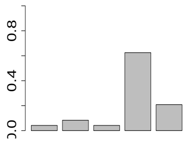
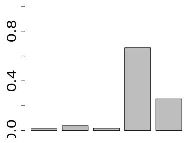
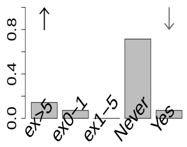
	tabaco (tobacco)
M (24)	
YW (51)	
WM (14)	

Table 5.20: Tobacco characteristics ∼ Baseline (Pre)

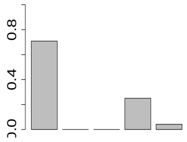
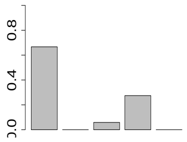
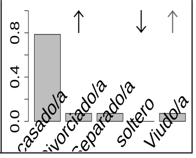
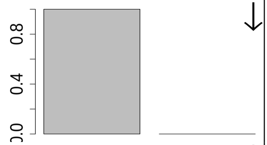
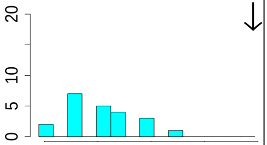
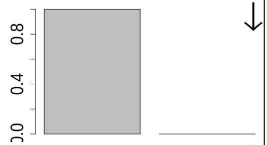
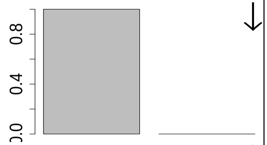
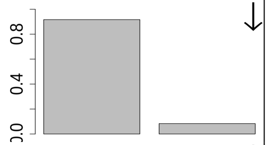
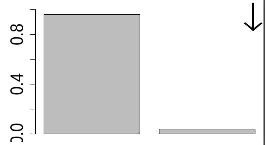
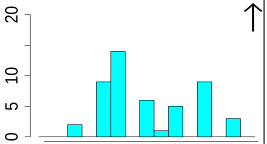
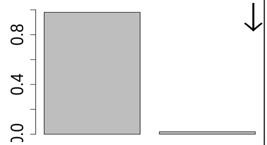
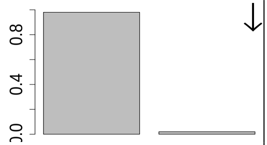
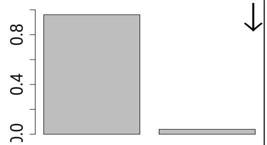
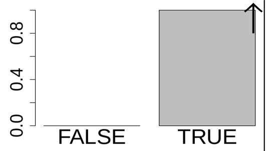
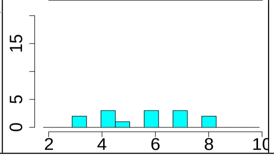
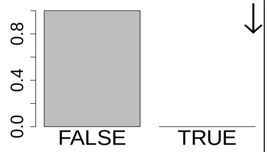
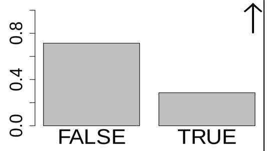
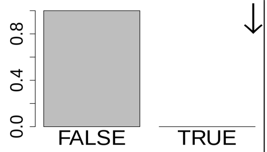
	est_civi (civilState)
M (24)	
YW (51)	
WM (14)	

Table 5.21: Sociodemographic characteristics ∼ Baseline (Pre)

	menopausia_bin (menopause)	pertensa (stressful)	depre (depression)	cancer (cancer)	fractura (bone fracture)
M (24)					
YW (51)					
WM (14)					

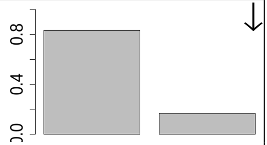
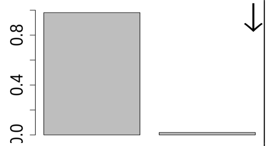
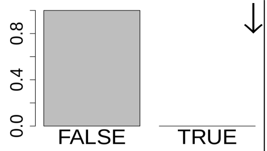
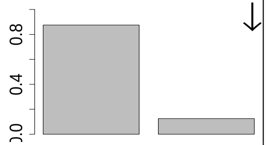
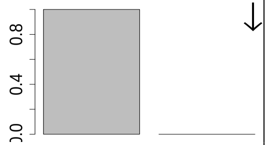
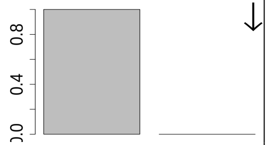
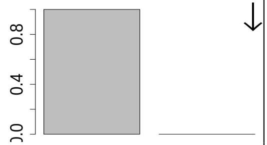
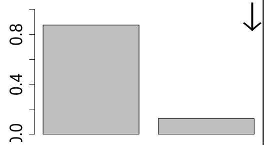
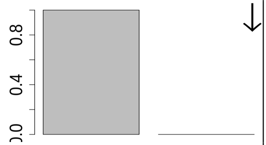
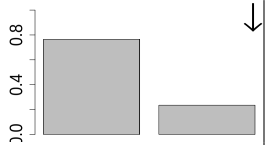
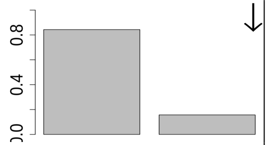
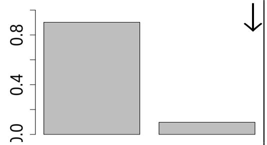
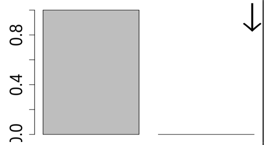
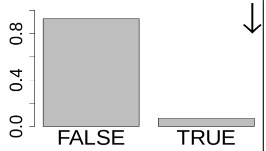
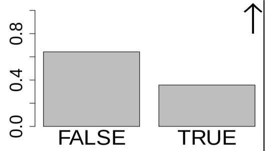
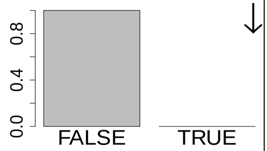
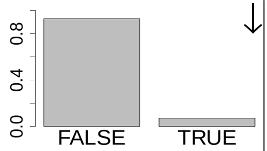
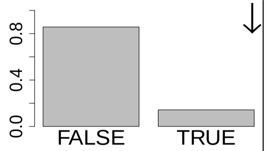
	disnea (dyspnea)
M (24)	
YW (51)	
WM (14)	

Table 5.22: Menopause & Diseases characteristics ∼ Baseline (Pre)

	maspirina1 (aspirin)	mdolor1 (NSAID)	mansiedad1 (anxiolytic)	mvitMin1 (vitamin or minerals)	mtension1 (heart)
M (24)					
YW (51)					
WM (14)					

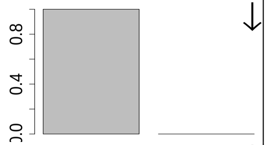
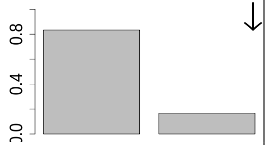
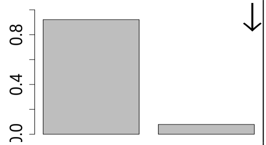
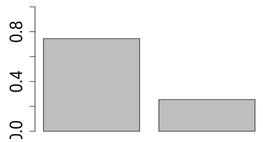
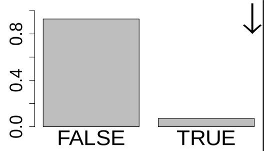
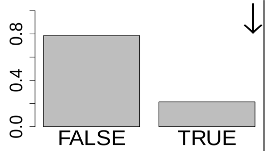
	mhormo1 (hormon)	motros1 (other)
M (24)		
YW (51)		
WM (14)		

Table 5.23: Drugs characteristics ~ Baseline (Pre)

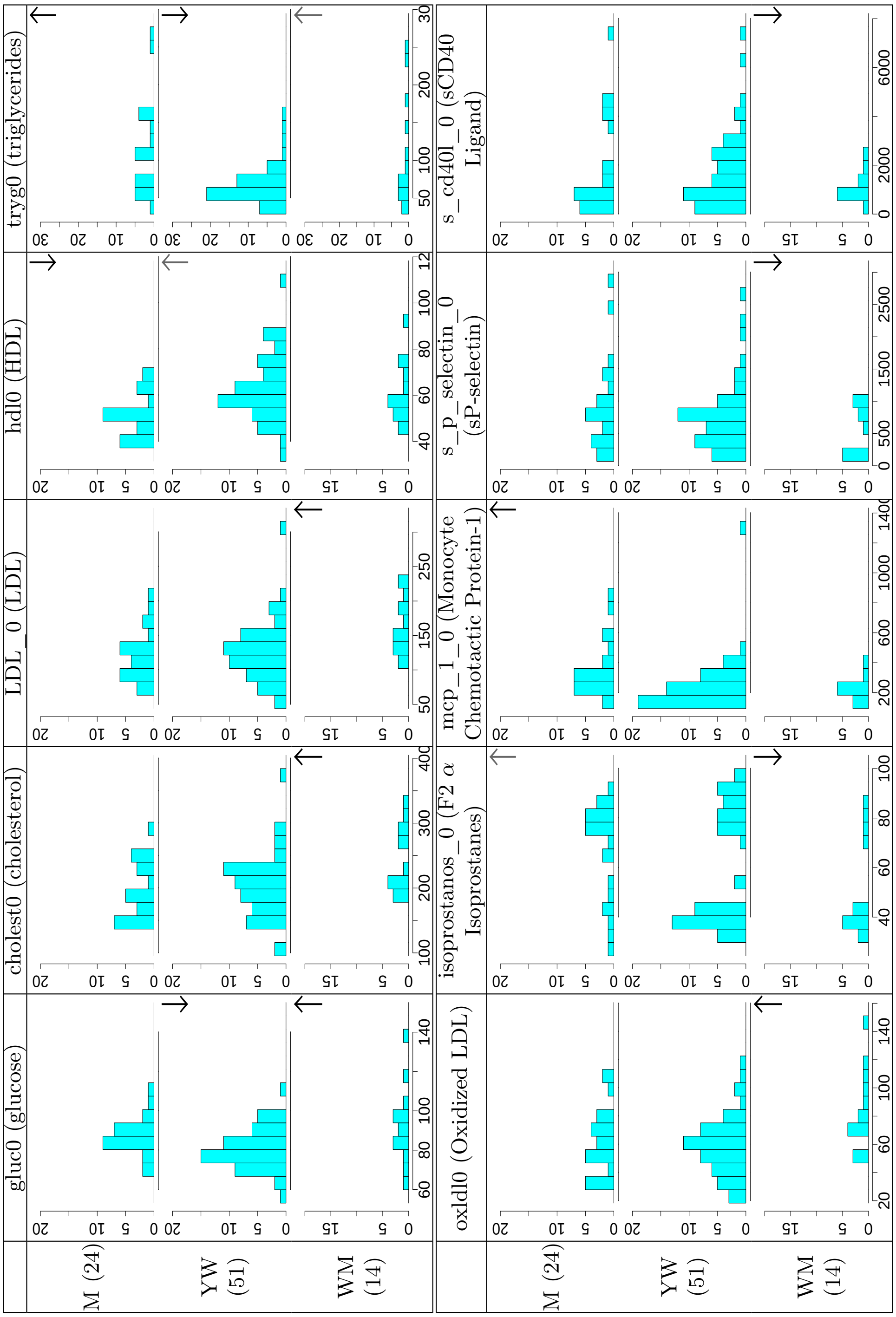


Table 5.24: Biomarkers characteristics ∼ Baseline (Pre) (a)

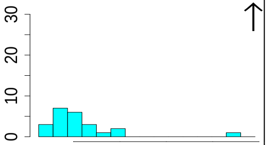
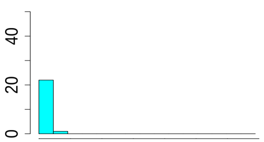
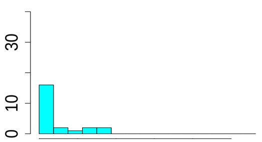
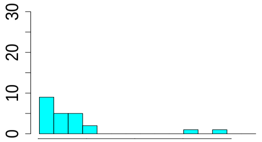
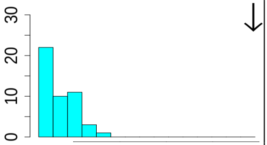
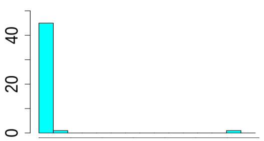
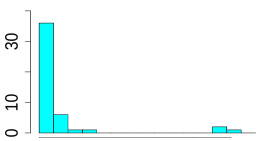
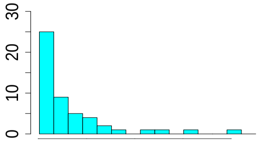
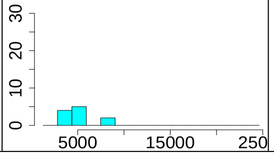
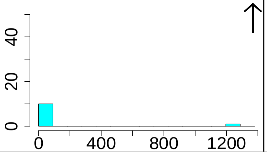
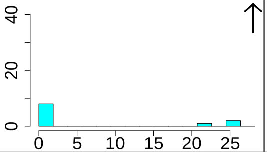
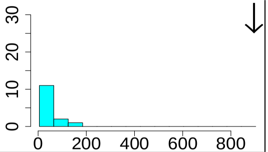
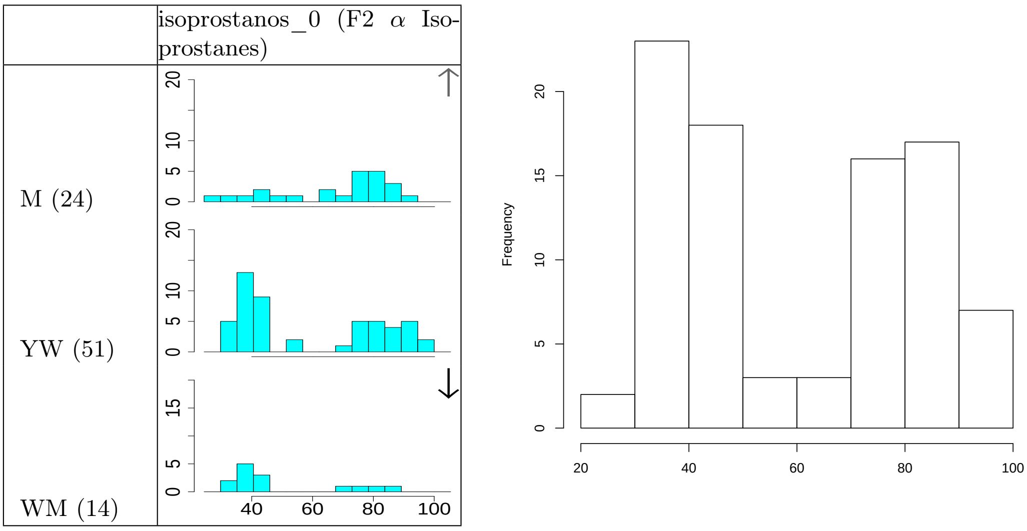
	t_pa_0 (Tissue Plasminogen Activator)	il_8_0 (Interleukin 8)	il_10_0 (Interleukin 10)	tyru0 (Tyrosol)
M (24)				
YW (51)				
WM (14)				

Table 5.25: Biomarkers characteristics ∼ Baseline (Pre) (b)

In the CPG 5.10a in Figure 5.10, the attribute *isoprostanos_0* (F2 α Isoprostanes) shows a bimodal behavior in cluster C_2. In Figure 5.10b, the attribute is shown and it is possible to identify that this behavior is also governed by the bimodal behavior of the entire attribute.



(a) Class Panel Graphs

(b) Histogram

Figure 5.10: F2 α Isoprostanes Biomarker

In Appendix B.1.1, the behavior of the final descriptors are shown after applying the NCI-IMS methodology and treating all the possible consistency problems with the nested partition P_0 as it will be seen in Section 5.4.5.

The resulting profiles are transcribed in natural language from the automatic profiles on the basis of the attributes behaving different in the several classes according to the NCI-IMS Cluster Interpretation Methodology.

Men (M): Group mainly composed by men that are taller and heavier (weight, BMI, waist). With higher blood pressure (systolic, diastolic) and lower heart rate. No cases of menopause. They consider themselves less stressful. Higher incidence of dyspnea than other groups and no other incidences of diseases. Low consumption of drugs. Lower levels of HDL and higher of triglycerides, F2 α Isoprostanes, Monocyte Chemotactic Protein-1 and Tissue Plasminogen Activator. Higher level of glucose (W) and lower of cholesterol (W) This group is renamed as "Men" (M).

Younger Women (YW): Group mainly composed by woman that are younger. They are smaller and thinner (weight, BMI, waist) with lower blood pressure (systolic, diastolic). Two cases of menopause. They consider themselves more stressful. No incidences of other diseases. Low consumption of drugs but they take more anxiolytics than general population. Lower levels of glucose, triglycerides and Tissue Plasminogen Activator and higher levels of HDL. Oxldl levels are lower (W). This group is renamed as "Young Women" (YW).

Women with Menopause (WM): Women group that are older. They are smaller and slightly thinner (weight) with higher blood pressure (systolic, diastolic) and heart rate. Higher proportion of long-term ex-smokers and lower of smokers. Mostly are married. All have menopause. Higher incidence of cancer and no incidences of other diseases. Higher consumption of painkillers (NSAID) but lower consumption, in general, of other drugs. Higher levels of glucose, cholesterol, LDL, triglycerides, Oxidized LDL, Interleukin 8 and Inerleukin 10. Lower levels of F2 α Isoprostanes, Monocyte Chemotactic Protein-1 (W), sP-selectin, sCD40 Ligand and Tyrosol. This group is renamed as "Women with menopause" (WM).

5.4.4 Characterization and Interpretation of the partition P_{H_o}

In this section, the attributes used for clustering are analyzed against the 3 resulting clusters. Although, in the clustering are included 4 of the 10 total physical activity attributes, the 10 attributes are used for the interpretation. As in previous Section 5.4.1, the resulting profiles are described on the basis of the attributes behaving different in the several classes resulting from the Cluster Interpretation Methodology CI-IMS (proposed in Section 3.6).

The resulting descriptor-powers and senses from the CI-IMS methodology for all classes are presented in Appendix B.2.1. After analyzing the consistency with the nested partition P_o, the profiles are modified, the resulting descriptors are drawn on the Class Panel Graphs included in Appendix B.2.2. In Appendix B.2.3 are included the automatic profiles as a result of the NCI-IMS methodology after analyzing the possible inconsistencies with the nested partition P_o (see Section 5.4.5).

The corresponding profiles with natural language transcription are the following:

The general sample is characterized for using olive oil as a main fat for cooking, consuming preferably white meat over other types of meat and preparing more than two times per week homemade tomato sauce.

The overall consumption of butter, gas drinks, wine and legumes is low. Although more than half of the individuals in all the classes have a low consumption, there are variations between different classes.

The resulting profiles are the following:

Based on White Meat (WMbased): Diet based on white meat and might include vegetables or fish. They have low consumption of fruit, red meat and legumes. It is a low caloric diet.

White Meat based with Sugars (WMwSugars): Diet rich in white meat, fruit and commercial bakery. May Occasionally include more nuts but not red meat neither vegetables. They practice more moderate and total exercise than the general sample.

Unhealthy (UH): Basically, they eat white meat with a lack of vegetables, wine, fish and nuts. Either they do not consume legumes neither olive oil or they consume red meat and gas drinks. They practice less moderate and total exercise in general.

5.4.5 Characterization and Interpretation of the Cross Partition P_o

In this section, the attributes involved in both baseline characteristic (C_o) and habits characteristics (H_o) are analyzed against the 8 crossed clusters.

Since this partition is built by crossing two partitions (P_{C_o} and P_{H_o}), this partition P_o is nested to both partitions P_{C_o} and P_{H_o}. For that reason, the methodology will be applied to this partition using P_{C_o} as the superclass for those attributes used to build this partition and P_{H_o} as the superclass for the corresponding attributes of habits. In this way, the characteristics of both P_{C_o} and P_{H_o} are used to build the description of P_o.

The Clustering Interpretation methodology NCI-IMS proposed in Section 3.9 is applied over the partition P_o and P_{C_o} with the attributes of the thematic block C. First, the interpretation of both partitions are obtaining using the CI-IMS methodology. The parameters are $\delta = 0.25$ admit till a 25% of contamination to consider a class described by a value, $\alpha = 0.01$ among the habitual values in hypothesis tests, $\varepsilon_1 = 0.3$ and $\varepsilon_2 = 0.2$ to make the Sensitive Analysis with $\{0.5n, 0.7n, 1, 1.3n, 1.5n\}$ and $r = \overline{R}$ for obtaining all possible descriptors for the Sensitive Analysis. Then, the consistency between both interpretations for each descriptor is assessed as it is described in Section 3.8. An analogous process is used with the partition P_{H_o} for the attributes of the thematic block H.

In Appendix B.3.1, the tables containing the descriptor-power and sense from the CI-IMS methodology are shown. The result of the NCI-IMS methodology for partition P_o are drawn on the Class Panel Graphs in Appendix B.3.2. The consistency of the descriptors is analyzed with partition P_o and partition P_{C_o} for the C attributes and with partition P_{H_o} for H attributes.

In Table 5.26, the number of descriptor-power for partitions P_{C_o} and P_o and attributes of thematic block C are shown. Only two attributes requires special attention. The descriptors power of s_vcam_1_0 (Soluble cell adhesion molecules-1) in class YW and ohtyru0 (OHTyrosol) for class WM, that are uncertain according to Table 3.10 in Section 3.8. In both cases, the descriptor can be omitted in the profile because it is not robust enough neither in the superclass nor in the corresponding subclasses.

In Table 5.27 are shown the number of descriptors power for partitions P_{H_o} and P_o and attributes of thematic block H. The descriptors power of the attribute ge2_3_5s_1

Table 5.26: Descriptors Power for partition P_{C_o} and P_o and attributes of thematic block C

	SuperClass							
SubClasses	**Robust Non-descriptor**	**Moderate Non-descriptor**	**Weak Non-descriptor**	**Weak Descriptor**	**Moderate Descriptor**	**Robust Descriptor**	**Basic Descriptor**	Sum
Robust Non-descriptor	20	1	0	0	0	0	0	21
Moderate Non-descriptor	2	0	3	0	0	0	0	5
Weak Non-descriptor	0	1	1	0	0	0	0	2
Weak Descriptor	9	0	1	0	0	0	0	10
Moderate Descriptor	9	0	1	0	1	0	0	11
Robust Descriptor	21	0	3	5	8	38	0	75
Basic Descriptor	1	0	0	0	0	1	36	38
Sum	62	2	9	5	9	39	36	162

(*lightWeek*) is *Weak* both for class UH and for the representative subclass. As in previous cases, this descriptor is not robust enough to describe neither the superclass nor the representative subclass.

Table 5.27: Descriptors Power for partition P_{H_o} and P_o and attributes of thematic block H

	SuperClass							
SubClasses	**Robust Non-descriptor**	**Moderate Non-descriptor**	**Weak Non-descriptor**	**Weak Descriptor**	**Moderate Descriptor**	**Robust Descriptor**	**Basic Descriptor**	Sum
Robust Non-descriptor	2	0	0	0	0	0	0	2
Moderate Non-descriptor	0	0	0	0	0	0	0	0
Weak Non-descriptor	1	0	0	0	0	0	0	1
Weak Descriptor	4	0	0	1	0	0	0	5
Moderate Descriptor	1	0	0	3	0	0	0	4
Robust Descriptor	15	0	0	0	1	7	0	23
Basic Descriptor	2	0	2	1	0	7	25	37
Sum	25	0	2	5	1	14	25	72

In Section 5.4.3 the interpretation of P_{C_o} is detailed using the CI-IMS methodology, and the final profiles of the classes after applying the modifications of the NCI-IMS methodology are shown. In Section 5.4.4, the final profiles of P_{H_o} after applying the NCI-IMS methodology are described.

Finally, the profiles for the partition P_o are automatically created in Appendix B.3.3

and a natural language transcription is the following.

M-WMbased: Men group, taller and heavier (weight, waist). Their blood pressure is higher (systolic, diastolic) and lower heart rate. Most are married. No menopause cases and they consider themselves less stressful. No incidence of other diseases. Low consumption of drugs. Lower level of HDL, Interleukin 6 and Interleukin 10. Higher levels of F2α-Isoprostanes, Monocyte Chemotactic Protein-1, Tyrosol, OHTyrosol, MOH Tyrosol. With a diet White Meat based. They consume more butter than general sample and less gas drinks and commercial bakery. They practice more exercise, specially intense, but they do less homework.

M-WMwSugars: Men group that are older, taller and heavier (weight, BMI, waist). With higher systolic and diastolic pressure. Most have never smoked. Most are married. No menopause cases. They consider themselves less stressful. Higher incidence of bone fracture and dyspnea. Low drugs consumption but higher intake of "other" drugs. Lower levels of LDL, HDL, sCD40 Ligand, C-Reactive Protein and Tyrosol. Higher levels of glucose, triglycerides, Monocyte Chemotactic Protein-1, sP-selectin and Tissue Plasmingen Activator. With a White Meat based with sugars diet. They consume legumes and less red meat and they drink more wine and less gas drinks. They practice more moderate and light exercise and more exercise in general but less intense and homework.

M-UH: Younger men group that are taller and heavier (weight, BMI, waist). With higher systolic pressure and lower heart rate. Lower proportion of non-smokers. There are higher proportion of singles. No menopause cases. They consider themselves less stressful. No incidence of other diseases. They do not intake drugs, as the general population. Lower levels of cholesterol, LDL, HDL and 8-Oxoguanine. Higher levels of triglycerides, F2α-Isoprostanes, sP-selectin, sCD40 Ligand and Tissue Plasminogen Activator. With the Unhealthy diet. They consume more red meat and gas drink and less butter than the general sample. They practice less exercise in general.

WY-WMbased: Group composed mainly by women. They are younger. They are thinner (weight, BMI, waist). Their systolic and diastolic pressure is lower. Higher proportion of smokers than other groups. No menopause in general. They consider themselves more stressful. No incidence of other diseases. They do not intake drugs as general population but they take more painkillers. Lower levels of glucose, triglycerides, Monocyte Chemotactic Protein-1 and higher of HDL and C-Reactive Protein. With a White Meat based diet, they eat more vegetables and fish. They consume little butter and wine.

WY-WMwSugars: Woman group which are smaller and thinner (weight, BMI, waist). Lower systolic pressure. Contains the unique depressed person and no incidence of other diseases. They do not intake drugs as general population. Lower levels of cholesterol, LDL, triglycerides, Oxidized LDL, C-Reactive Protein and Tissue Plasminogen Activator. Higher levels of Interferon-γ, Tumor necrosis factor-α, OHTyrosol and MOHTyrosol. With a diet White Meat with sugars. They consume more nuts and less red meat and legumes. They practice more intense exercise and more exercise than in general population.

WY-UH: Younger women group that are smaller and with lower systolic and diastolic pressure. No menopause cases. They are more stressful. No incidences of other diseases. Low consumption of drugs but slightly higher of "other" drugs. Lower levels of glucose, cholesterol, LDL, triglycerides, Oxidized LDL, Tissue Plasminogen Activator and Interleukin 10. Higher levels of F2α-Isoprostanes, 8-Oxoguanine, Tumor necrosis factor-α, Soluble cell adhesion molecules-1, Interleukin 6 and Tyrosol. With an Unhealthy diet and they consume less olive oil, gas drinks and legumes and use more butter. They do more homework and less exercise in general.

WM-WMbased: Older women which are shorter and thinner (weight) they have higher diastolic pressure and heart rate. Most are married. All have menopause. They consider themselves more stressful. They have higher incidence of cancer with no other incidences of diseases. They do not intake main drugs, as general population, but they take more "other" drugs. Higher levels of glucose, cholesterol, LDL, HDL, Oxidized LDL, C-Reactive Protein, Interleukin 8 and Interleukin 10. Lower levels of F2α-Isoprostanes, sP-selectin sCD40 Ligand and Tyrosol. With a White Meat based diet and rich in vegetables and poor in butter, wine,commercial bakery and nuts. They do more light exercise but practice less in general.

WM-WMwsugars: Older women group which are shorter and heavier (BMI, waist). They have higher blood pressure and heart rate. Lower proportion of smokers. Lower proportion of singles. All have menopause. No incidence of other diseases. They take more painkillers (NSAID) and not other drugs. They have higher levels of glucose, cholesterol, LDL, triglycerides, Oxidized LDL and Interleukin 10. Lower levels of HDL, Monocyte Chemotactic Protein-1, sP-selectin, sCD40 Ligand, C-Reactive Protein and Tyrosol. With a White Meat base with sugars diet. They consume less vegetables, red meat, gas drinks and wine. They eat more nuts. They do less intense exercise and more homework.

5.5 Profiling Final State of Individuals

In this section, the goal is to find different profiles of subjects depending on the final (post) characteristics of the individuals and their living habits at the end of the study.

Using the same methodological approach of the Integrative Multiview Clustering, the profiles after the intervention are built.

The data used for this clustering is composed by the same 65 attributes but some of them measured after the intervention. Therefore, the final classes are built using the intrinsic attribute matrix (χ) that are those that do not change over the study, such as the *gender* and thus, there is only one measure. The matrix defined as $\mathcal{Y}_f$ contains the second measures of $\mathcal{Y}$.

For the Integrative Multiview Clustering the attributes are split in the same two thematic blocks defined in Section 5.3. The corresponding blocks with the post measures are: C_f and H_f.

Therefore, the resulting partition is a crossing of two partitions: state of participants in 3 classes (M, W, MW) and habits in 4 classes (H↑OO, HB, BasicwBak, ProtoCaloric). Finally, a partition P_f of 12 classes is obtained.

5.5.1 Clustering of the Final Characteristics after the Intervention (C_f)

From the 47 attributes of the thematic block C_f, 7 are intrinsic attributes ($\mathcal{X}$) and 40 are the second measures of attributes $\mathcal{Y}$ (see details in Section 5.3).

Figure 5.11 depicts the dendrogram of applying hierarchical clustering with Ward's method and using the Gower's metrics to this dataset.

Figure 5.12a shows the histogram of level indexes of the dendrogram (Figure 5.12a). From this picture, it possible to determine a convenient cut in 2 or 3 clusters according to max leaps in the graphic.

Finally, we selected 3 clusters: $P_{C_f} = \{FC_1, FC_2, FC_3\}$. Figure 5.12b shows the cut in these 3 clusters over the dendrogram. Table 5.28 shows the classes sizes and the individuals of each class.

Table 5.28: Partition P_{C_f} basic information

Cluster	Size	Individuals
FC_1	26	$\{i_1, i_{12}, i_{13}, i_{14}, i_{15}, i_{18}, i_{20}, i_{30}, i_{31}, i_{39}, i_{40}, i_{41}, i_{48}, i_{53}, i_{54}, i_{59}, i_{62}, i_{64}, i_{66}, i_{67}, i_{71}, i_{79}, i_{81}, i_{85}, i_{88}, i_{89}\}$
FC_2	48	$\{i_2, i_3, i_4, i_5, i_6, i_7, i_8, i_9, i_{10}, i_{19}, i_{22}, i_{23}, i_{24}, i_{27}, i_{28}, i_{29}, i_{32}, i_{33}, i_{34}, i_{35}, i_{36}, i_{37}, i_{42}, i_{43}, i_{44}, i_{45}, i_{46}, i_{47}, i_{49}, i_{50}, i_{51}, i_{52}, i_{55}, i_{56}, i_{57}, i_{58}, i_{60}, i_{61}, i_{63}, i_{65}, i_{68}, i_{69}, i_{70}, i_{72}, i_{76}, i_{80}, i_{86}, i_{87}\}$
FC_3	15	$\{i_{11}, i_{16}, i_{17}, i_{21}, i_{25}, i_{26}, i_{38}, i_{73}, i_{74}, i_{75}, i_{77}, i_{78}, i_{82}, i_{83}, i_{84}\}$

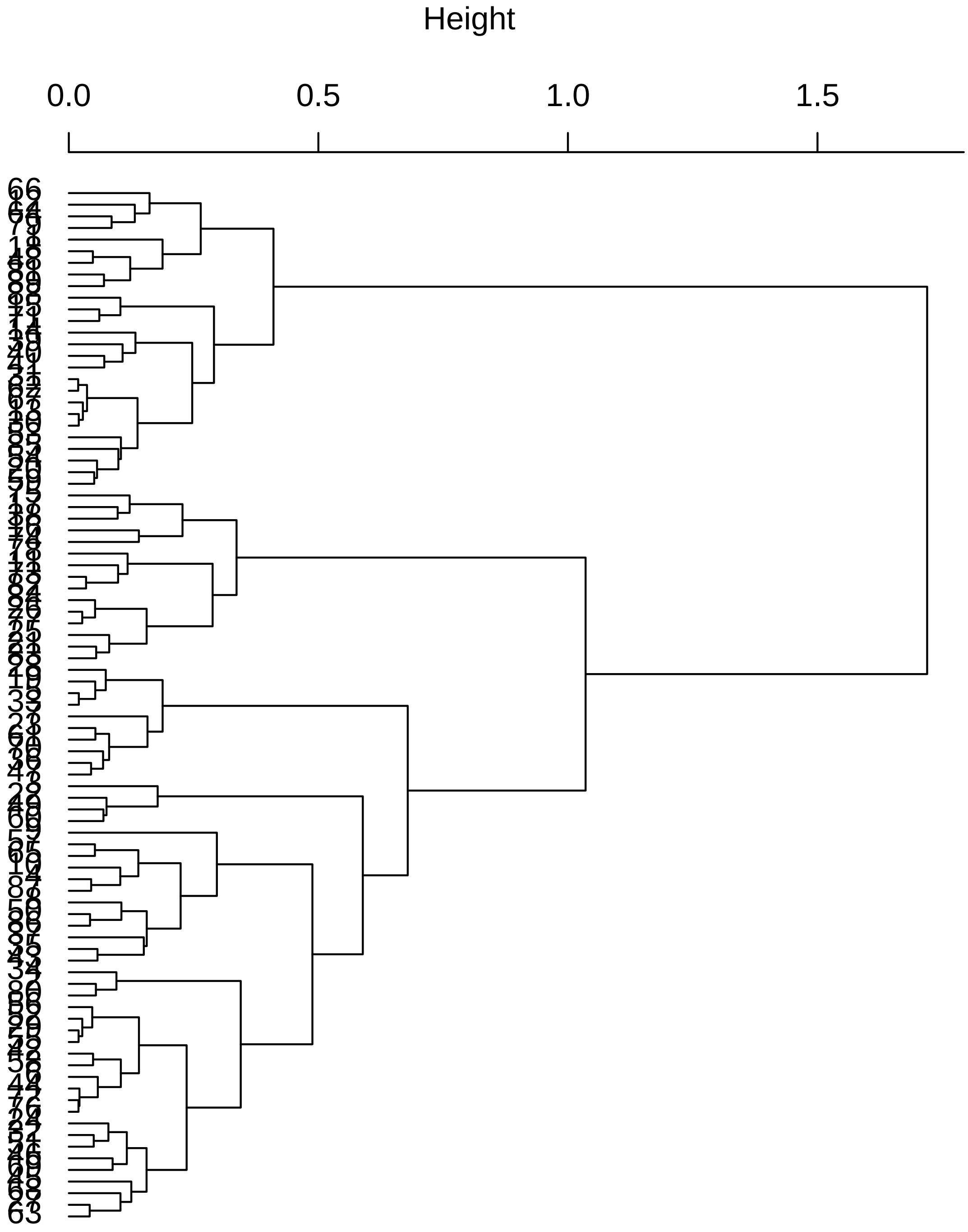
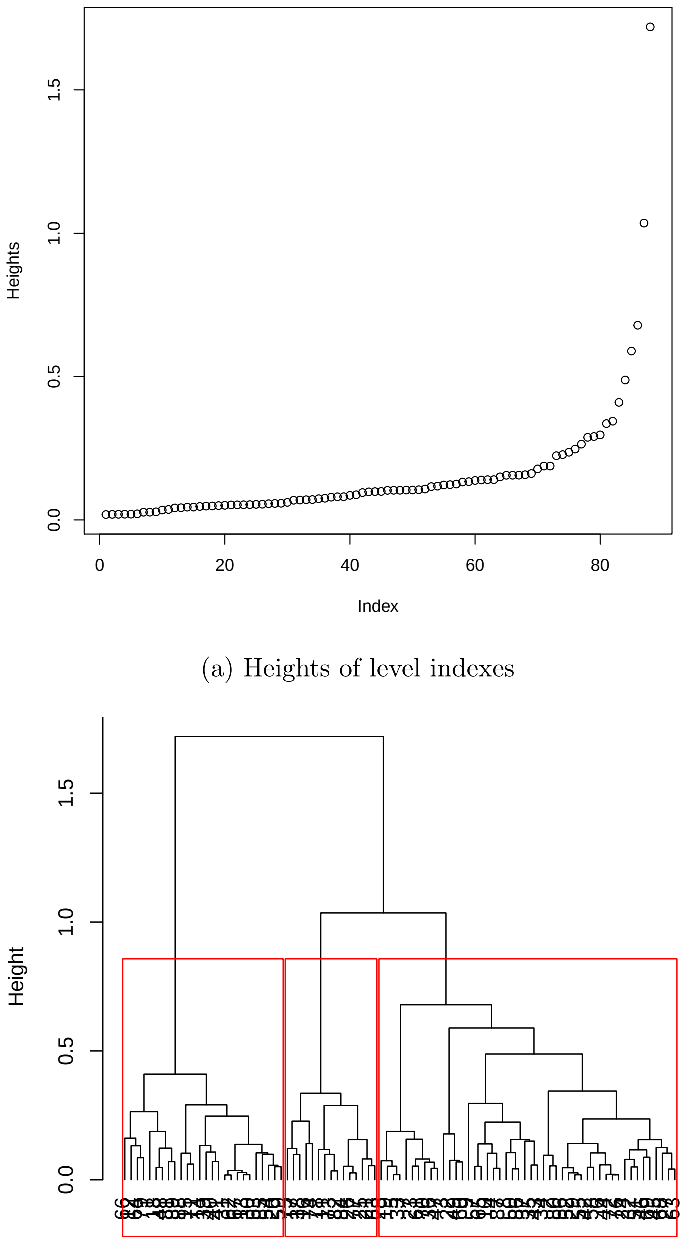


Figure 5.11: Hierarchical clustering with Ward's Method for final characteristics



(a) Heights of level indexes

(b) Dendrogram cut in 3 clusters

Figure 5.12: Hierarchical clustering with Ward's Method for Final characteristics

Although, the best partition is in 2 clusters, in Table 5.29 cutting in 3 groups is compared with the initial clusters (P_{C_o}) found in Section 5.4.1. The 82 of 89 match the first classification (see Table 5.29), for this reason 3 is selected in order to compare the profiles.

Table 5.29: Comparison of classes of P_{C_o} and P_{C_f}

	P_{C_o}		
P_{C_f}	M	YW	WM
FC_1	24	2	0
FC_2	0	48	0
FC_3	0	1	14

As in the initial state, gender and menopause are the main attributes that define these clusters. In Table 5.30 are the contingency tables of the *gender* and *menopause* attributes. Both tables show that the resulting clusters can be defined by these attributes: Men, Women without menopause, Women with menopause.

Table 5.30: Contingency table of Gender and Menopause of classes of P_{C_f}

	Gender			Menopause	
	Man	Woman		FALSE	TRUE
FC_1	25	1	FC_1	26	0
FC_2	0	48	FC_2	47	1
FC_3	0	15	FC_3	0	15

After the characterization of the partition in Section 5.5.4, the classes are renamed as: Men (M), Women (W), Women with Menopause (WM).

5.5.2 Clustering of Habits after the Intervention (H_f)

The second clustering is applied over the attributes referred to the diet and physical activity at the end of the study. These attributes belong to the diet (p14-item) and physical activity questionnaire of the second visit in the study.

The dendrogram of the resulting hierarchical clustering with Ward's method and using Gower's metrics is depicted in Figure 5.13. Figure 5.14a is the histogram of level indexes of the dendrogram (Figure 5.13). From this picture, it possible to determine a convenient cut in 4 clusters according to max leaps in the graphic.

Finally, the 4 clusters: $P_{H_f} = \{FH_1, FH_2, FH_3, FH_4\}$. Figure 5.14b shows the cut in these 4 clusters over the dendrogram. Table 5.31 shows the sizes and the individuals of each class.

In Table 5.32 is shown the comparison between both classes of habits. The habits at the beginning of the study (P_{H_o}) and at the end (P_{H_o}) are quite different and the individuals are not grouped in the same clusters.

Table 5.31: Information of Partition P_{H_f}

Cluster	Size	Individuals
FH_1	26	$\{ i_1, i_5, i_6, i_{12}, i_{13}, i_{16}, i_{22}, i_{23}, i_{30}, i_{33}, i_{42}, i_{45}, i_{52}, i_{54}, i_{56}, i_{59}, i_{63}, i_{64}, i_{68}, i_{69}, i_{78}, i_{80}, i_{83}, i_{84}, i_{85}, i_{88}\}$
FH_2	17	$\{ i_3, i_4, i_8, i_{15}, i_{18}, i_{20}, i_{21}, i_{24}, i_{31}, i_{32}, i_{34}, i_{35}, i_{38}, i_{60}, i_{62}, i_{86}, i_{89}\}$
FH_3	19	$\{i_7, i_9, i_{10}, i_{11}, i_{19}, i_{29}, i_{40}, i_{41}, i_{46}, i_{47}, i_{49}, i_{55}, i_{57}, i_{61}, i_{65}, i_{72}, i_{73}, i_{75}, i_{77}\}$
FH_4	27	$\{ i_2, i_{14}, i_{17}, i_{25}, i_{26}, i_{27}, i_{28}, i_{36}, i_{37}, i_{39}, i_{43}, i_{44}, i_{48}, i_{50}, i_{51}, i_{53}, i_{58}, i_{66}, i_{67}, i_{70}, i_{71}, i_{74}, i_{76}, i_{79}, i_{81}, i_{82}, i_{87}\}$

	P_{H_o}		
P_{H_f}	WMBased	WMwSugars	UH
FH_1	14	5	7
FH_2	2	18	7
FH_3	13	0	4
FH_4	14	5	0

Table 5.32: Comparison between Initial Habits and Final Habits

After the characterization of the partition in Section 5.5.5, the classes are renamed as: Healthy with Olive Oil (H↑OO), Healthy Basic (HB), Healthy with High consumption of Commercial bakery(BasicwBak) and Proteic Caloric (ProtoCaloric).

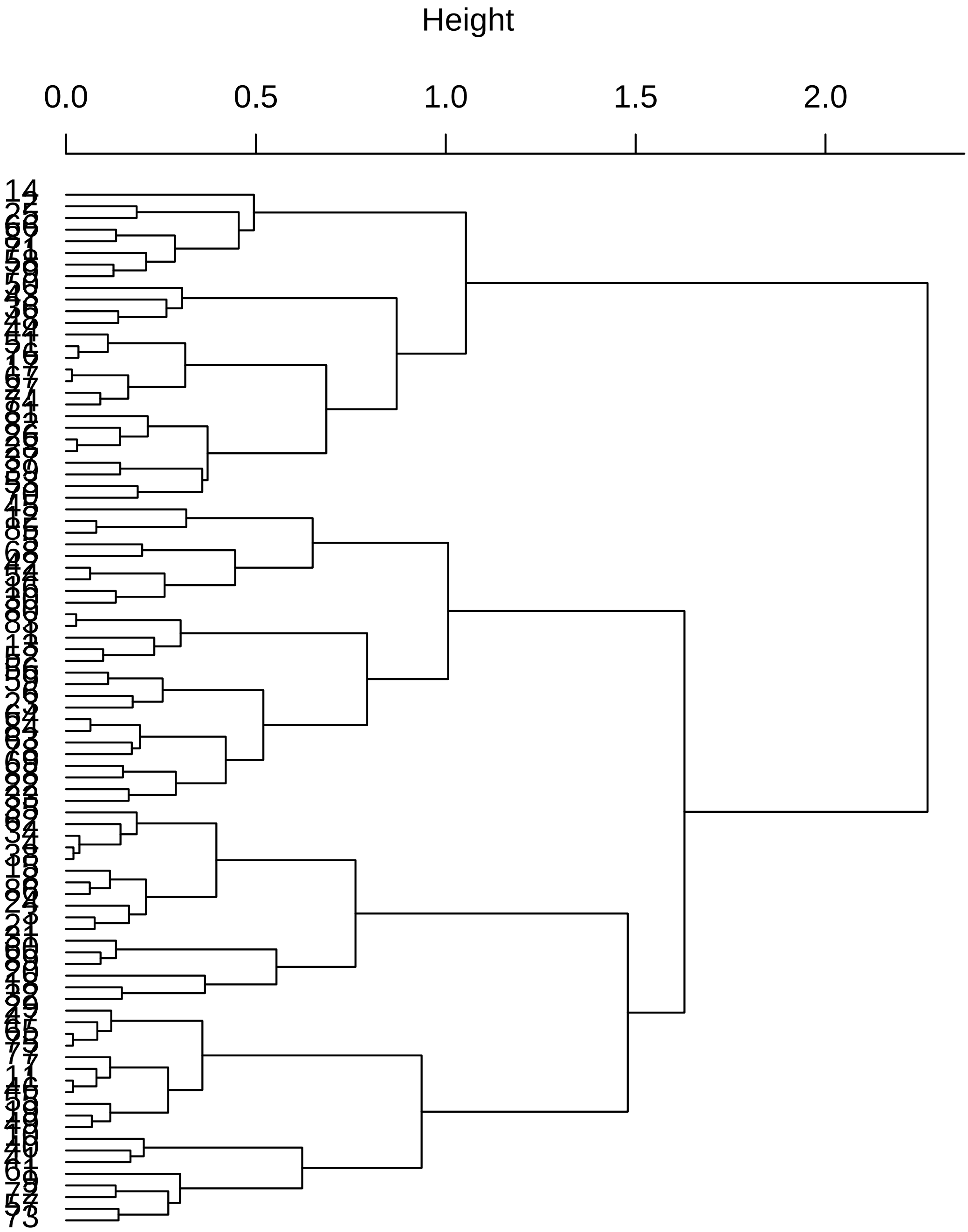
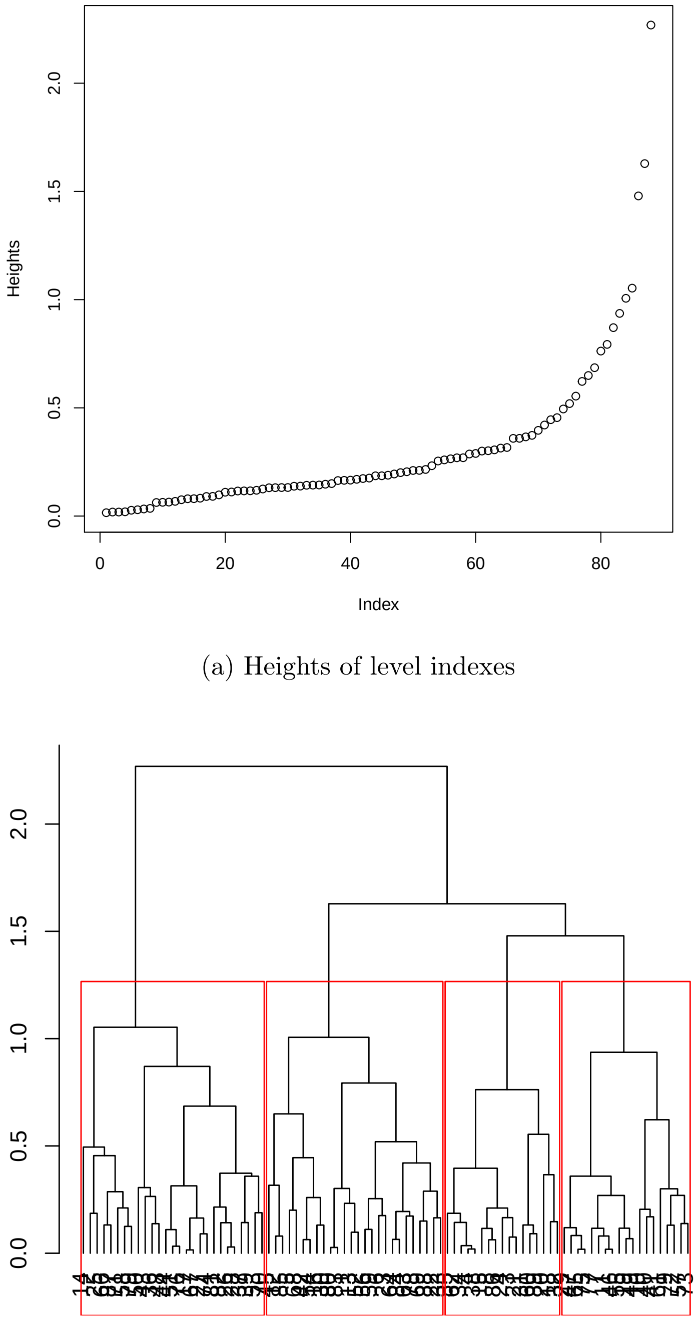


Figure 5.13: Hierarchical clustering with Ward's Method for final habits characteristics



(a) Heights of level indexes

(b) Dendrogram cut in 4 clusters

Figure 5.14: Hierarchical clustering with Ward's Method for final characteristics

5.5.3 Building the Cross Clustering with Final Characteristics and Final Habits

In next Table 5.15, the clusters of both final characteristics ($P_{C_f} = \{$M, W, WM$\}$) and habits ($P_{H_f} = \{$H↑OO, HB, BasicwBak, ProtoCaloric$\}$) clusterings are crossed ($P_f = P_{C_f} \times P_{H_f}$).

Table 5.33: Crossing Final Characteristics and Final Habits

	P_{C_f}		
P_{H_f}	M	W	WM
H↑OO	9	13	4
HB	6	9	2
BasicwBak	2	13	4
ProtoCaloric	9	13	5

Therefore, P_f is composed by the following classes:

1. M-H↑OO
2. M-HB
3. M-BasicwBak
4. M-ProtoCaloric
5. W-H↑OO
6. W-HB
7. W-BasicwBak
8. W-ProtoCaloric
9. WM-H↑OO
10. WM-HB
11. WM-BasicwBak
12. WM-ProtoCaloric

5.5.4 Characterization and Interpretation of the partition P_{C_f}

The Cluster Interpretation Methodology CI-IMS has been applied over the resulting partition P_{C_f}. In this case, mostly attributes are second measures, for that reason the *Test-Value* used in the methodology is the dynamic generalized version defined in Table 3.1 in Section 3.5.Therefore, in order to compare both partitions, the interpretation of P_{C_f} is assessed comparing the values of each class against the baseline measures. Thus, the difference is the assessment of the statistical test. For the numerical attributes, the mean of the global attribute at the beginning of the intervention is compared against the mean of the class at the end of the intervention. For the categorical attributes is analogous: the baseline global proportion is compared against the proportion of the class at the end of the intervention. This dynamic generalized Test-Value is used, also, in the interpretation of P_{H_f} in Section 5.5.5 and in the interpretation of P_f (Section 5.5.6).

The resulting descriptor-powers with the corresponding senses from the CI-IMS methodology for all classes are included in Appendix B.4.1. Class Panel Graphs are included in Appendix B.4.2. The resulting descriptors after applying the NCI-IMS methodology using the nested partition P_f are drawn on the CPG.

The following profiles correspond to natural language of the automatic profiles generated by the NCI-IMS methodology which are included in Appendix B.4.3 :

Men (M): Group composed mainly by men. They are younger, taller and heavier (weight, BMI, waist) with higher blood pressure (systolic, diastolic) and lower heart rate. No cases of menopause. They consider themselves less stressful. No incidences of other diseases. Low consumption of drugs.

Lower levels of cholesterol, LDL, HDL and 8-Oxoguanine. Higher levels of triglycerides and sP-selectin. Diet rich in red meat, butter and homemade sauces. Poor in wine and commercial bakery. They do less homework. This class is renamed as Men (M)

Women (W): Women group that are younger, smaller and thinner (weight, BMI, waist). With lower blood pressure (diastolic, systolic) and heart rate. One case of menopause. They consider themselves more stressful. No incidences of other diseases. Low consumption of drugs but they take more anxiolytics than general population. Lower levels of glucose, triglycerides, F2 α Isoprostanes, 8-Oxoguanine and Tissue Plasminogen Activator; higher levels of Soluble cell adhesion molecules-1. Diet rich in vegetables and poor in red meat, butter, wine and legumes. This class is renamed as Women (W)

Women with menopause (WM): Women group that are older. They are smaller and slightly thinner (weight) with higher blood pressure (systolic, diastolic). Higher proportion of non-smokers. Most are married. All have menopause. Higher incidence of cancer and no incidences of other diseases. Higher consumption of hormones than other groups, low consumption of other drugs. Higher levels of cholesterol, LDL, triglycerides, Oxidized LDL, C-Reactive Protein and Tissue Plasminogen Activator. Lower levels of F2 α Isoprostanes, Monocyte Chemotactic Protein-1, sP-selectin, sCD40 Ligand, 8-Oxoguanine and Tyrosol. Diet poor in red meat, butter, gas drinks, wine, legumes and rich in fish. They practice less intense exercise and less exercise in general (W). This class is renamed as Women with menopause (WM)

5.5.5 Characterization and Interpretation of partition P_{H_f}

In this section, the attributes used for clustering are analyzed against the 4 resulting clusters. Although, in the clustering are included 4 of the 10 total physical activity attributes, the interpretation is made using all 10 attributes. As in previous Section 5.5.4, the profiles are the result of the Cluster Interpretation Methodology CI-IMS using the dynamic variant of the Generalized Test-Value (see Table 3.1) and NCI-IMS methodology to contrast the consistency with the nested partition P_f.

The resulting descriptor-powers with the corresponding sense from the CI-IMS methodology for all classes are included in Appendix B.5.1. The resulting descriptors of applying the NCI-IMS methodology are drawn on the corresponding Class Panel Graphs in Appendix B.5.2. In Appendix B.5.3 are included the automatic profiles as a result of the

NCI-IMS methodology. The following profiles correspond natural language transcriptions of these profiles:

The resulting 4 profiles are characterized for using olive oil as a main fat for cooking, consuming preferably white meat over other types of meat and preparing more than two times per week homemade tomato sauce.

In the overall consumption of butter, gas drinks, wine and legumes is low. Although more than half of the individuals in the sample have low consumption, there are variations between different classes.

Healthy with Olive Oil (H↑OO): Diet is rich in olive oil, wine, white meat and homemade sauces and poor in nuts. They eat either more vegetables or more fish.

Healthy Basic (HB): This is a low protein diet. It is poor red meat, gas drinks, wine, legumes, fish and commercial bakery and rich in white meat. They practice less moderate exercise.

Healthy with High consumption of Commercial bakery(BasicwBak): Diet rich in vegetables, white meat and commercial bakery. It is poor in butter, gas drinks and wine. They consume either more white meat or more red meat, legumes and fish. They practice more intense physical activity and total than general sample.

Proteic Caloric (ProtoCaloric): Diet is poor in vegetables and rich in legumes and commercial bakery. They consume either more red meat or more nuts or more fish and nuts.

5.5.6 Characterization and Interpretation of the Cross Partition P_f

In this section, the attributes involved in both blocks: Basic characteristics (C) and habits (H) are analyzed against the 12 crossed clusters. The attributes are graphically represented using class panel graphs in Appendix B.6.2.

The interpretation has been done following the described methodology NCI-IMS (see Section 3.9) like in the characterization of partition P_o in Section 5.4.5. The difference with the interpretation of P_o is that P_f is created using the second measures of most of the attributes. Therefore, as in previous interpretation of P_{C_f} in Section 5.5.1 and interpretation of P_{H_f} (Section 5.5.2), the underlaying Cluster Interpretation Methodology CI-IMS uses the dynamic variant of the Generalized Test-Value.

The resulting descriptor-powers with the corresponding senses from the CI-IMS methodology for all classes are shown in Appendix B.6.1. The results of the NCI-IMS methodology are drawn on the corresponding Class Panel Graphs in Appendix B.6.2. In Appendix B.6.3 the automatic profiles as the result of the NCI-IMS methodology are included.

The following profiles correspond to natural language transcriptions of the automatic profiles:

The general sample is characterized for using olive oil as a main fat for cooking, consuming preferably white meat over other types of meat and preparing more than two times per week homemade tomato sauce.

M-H↑OO: Men group taller and heavier (weight, BMI, waist). Their heart rate is lower. No menopause cases. They consider themselves less stressful and no incidences of other diseases. Low consumption of drugs. They have lower levels of HDL, F2α Isoprostanes. Higher levels of triglycerides, Oxidized-LDL, sP-Selectin, C-Reactive Protein, Tissue Plasminogen Activator and tyrosol. Levels of 8-Oxoguanine are lower than in the baseline state. With a healthy diet rich in olive oil and vegetables. They consume more red meat, gas drinks and wine and less butter. They practice more exercise than general sample and do less homework.

M-HB: Men group taller and heavier (weight, BMI, waist). With higher systolic and diastolic pressure and lower heart rate. Higher proportion of smokers than general sample. Most are married. No menopause cases. They consider themselves less stressful. No incidence of other diseases. Low intake of drugs. Lower levels of cholesterol, LDL, HDL, Oxidized-LDL, C-Reactive Protein and Tissue Plasminogen Activator. Levels of 8-Oxoguanine are lower than in the baseline state but they have higher values than others. Higher levels of F2α Isoprostanes. With a healthy basic diet. Higher proportion consume more butter than in other groups. They do less moderate exercise and more intense than general sample.

M-BasicwBak: Men group taller and heavier (weight, waist). With higher blood pressure (systolic, diastolic) and heart rate. All never smokes. All are married. No menopause cases. No other incidences of diseases. Low consumption of main drugs but they take more "other" drugs. Lower levels of cholesterol, LDL, HDL , triglycerides, Oxidized LDL, F2α Isoprostanes, C-Reactive Protein, Interleukin 6, Tyrosol, OH Tyrosol and MOH Tryrosol. Higher levels of Monocyte Chemotactic Protein-1 and sP-Selectin. Levels of 8-Oxoguanine are lower than in the baseline state and lower than other groups. With a basic diet rich in commercial bakery. They consume more red meat, legumes, fish and nuts. They do more intense resulting a higher practice in general but do less homework.

M-ProtoCaloric Men group younger, taller and heavier (weight, BMI, waist). With higher blood pressure (systolic, diastolic) and lower heart rate. Higher proportion of singles. No menopause cases. They consider themselves less stressful and no incidence of other diseases. They do not take drugs as general population. Lower levels of cholesterol, LDL, HDL, F2α Isoprostanes and Tyrosol. Higher levels of triglycerides, Monocyte Chemotactic Protein-1 and sP-Selectin. Levels of 8-Oxoguanine are lower than in the baseline state. With a proteic and caloric diet. Rich in red meat, gas drinks and legumes. They practice more moderate exercise and do less homework.

W-H↑OO: Women group slightly thinner (BMI, waist) with lower blood pressure (systolic, diastolic) and lower heart rate. Lower proportion of non-smokers than general sample. No menopause cases. They consider themselves more stressful. No incidences of other diseases. They do not take drugs as general population. Lower levels of triglycerides, Monocyte Chemotactic Protein-1, sCD40 Ligand, C-Reactive Protein and Tyrosol. Higher levels of Soluble cell adhesion molecules-1. Levels of 8-Oxoguanine are lower than in the baseline state. With a healthy diet rich in olive oil. They consume more gas drinks, wine, legumes and fish. They practice less exercise in general.

W-HB: Women group smaller and thinner (weight, waist) with lower systolic and diastolic pressure. Higher proportion of smokers. All are married. No menopause cases. They consider themselves more stressful. No incidence of other diseases. They take more painkillers, anxiolytics and "other" drugs than general sample. Higher levels of cholesterol, LDL, Oxidized LDL, C-Reactive Protein, Tumor necrosis factor-α, Soluble Cell Adhesion molecules-1, Interleukin 6 and Interleukin 10. Lower levels of glucose, sCD40 Ligand and Tissue Plasminogen tivator. Levels of 8-Oxoguanine are lower than in the baseline state. With a healthy basic diet and rich in vegetables. They practice less exercise.

W-BasicwBak: Women group smaller and thiner (weight, BMI, waist). With lower blood systolic and diastolic pressure and heart rate. Higher proportion of singles than general population. No menopause cases. They consider themselves more stressful. No incidence of other diseases. Low intake of drugs. Lower levels of glucose, tryglycerides, Monocyte Chemotactic Protein-1, sP-selectin, Tissue Plasminogen Activator, Interleukin 6 ad Interleukin 10. Levels of 8-Oxoguanine are lower than in the baseline state. With a basic diet rich in commercial bakery. They eat more white meat. They practice more exercise (intense and total) but do less homework.

W-ProtoCaloric: Women group slightly smaller and thinner (weight, BMI, waist) with lower systolic pressure and heart rate. Most of them are non-smokers. Only one case of menopause. No incidence of other diseases. They do not take drugs as general population with the exception of anxiolytics. Lower levels of glucose, cholesterol, LDL, triglycerides, Oxidized LDL, F2α Isoprostanes and Tissue Plasminogen Activator. Higher level of Interferon-γ, sP-secectin, sCD40 Ligand. Levels of 8-Oxoguanine are lower than in the baseline state. With a proteic and caloric diet. They consume more vegetal protein (legumes and nuts). Also, they consume fruit. They do more homework and intense exercise per year than the general sample.

WM-H↑OO: Older women that are smaller with higher systolic and diastolic pressure. Most never smoke. Most are married. All have menopause. Higher incidence of cancer and no incidence of other diseases. They do not take drugs as general population excepts for hormones. Higher levels of glucose, C-Reactive Protein and Tissue Plasminogen Activator. Lower levels of HDL, sP-selectin, sCD40 Ligand, Interleukin 6, Interleukin 10 and Tyrosol. Levels of 8-Oxoguanine are lower than in the baseline state. With a healthy diet rich in olive oil and fish. They consume more butter and wine than the general sample. They do less intense and total exercise.

WM-HB: Older women that are heavier (BMI, waist) with higher blood pressure (systolic, diastolic) and heart rate. Lower proportion of smokers. All are married. All have menopause. They consider themselves more stressful and have higher incidence of cancer. They take more tension drugs and low consumption of other drugs. Higher levels of glucose, cholesterol, LDL, HDL, trigycerides, oxidized LDL, C-Reactive Protein and Tissue Plasminogen Activator. Lower levels of F2α Isoprostanes, Monocyte Chemotactic Protein-1, sP-selectin, sCD40 Ligand, Interleukin 6, Interleukin 10 and Tyrosol. Levels of 8-Oxoguanine are lower than in the baseline state. With a healthy basic diet rich in vegetables. They practice less exercise than general sample.

WM-BasicwBak: Older women smaller and thinner (weight). With lower systolic pressure and heart rate. All never smoke. All are married. All have menopause. They have higher incidence of cancer and dyspnea. They take more painkillers (NSAID) and not other drugs. Higher levels of cholesterol, LDL, HDL, triglycerides, oxidized LDL, C-Reactive Protein, Interleukin 6 and Interleukin 10. Lower levels of F2 α Isoprostanes, sP-selectin, sCD40 Ligand, Tissue Plasminogen Activator, Tyrosol, OhTyrosol and MOHTyrosol. Levels of 8-Oxoguanine are lower than in the baseline state. With a basic diet rich in commercial bakery. They eat more white meat. They practice less exercise and do more homework.

WM-ProtoCaloric: Older women that are smaller and heavier (BMI, waist). Their blood pressure (systolic, diastolic) is higher. Most never smoke. Lower proportion of singles. All have menopause. They consider themselves less stressful. Higher incidence of cancer and no incidence of other diseases. They do not take main drugs, as general population, but they take more "other" drugs. Higher levels of cholesterol, LDL, triglycerides, Oxidized LDL and Tissue Plasminogen Activator. Lower levels of glucose, Monocyte Chemotactic Protein-1, sP-selectin, sCD40 Ligand, Interleukin 10 and Tyrosol. Levels of 8-Oxoguanine are lower than in the baseline state. With a proteic and caloric diet and rich in fruit. They consume proteins from legumes, fish and nuts. They do more light and moderate exercise per year than the general sample.

5.6 Pre-Post Trajectory Analysis

In this section, the initial state of individuals (P_o) is compared with the final state of the individuals (P_f) taking into account the intervention (T). This analysis is referred as *Pre-Post Trajectories Analysis*. The objective of this analysis is to observe how the individuals change their diet profiles during the study. Since this trajectory analysis takes into account the intervention, it can be analyzed whether individuals of one class behaves differently depending on the assigned intervention.

Figure 5.15 shows all the trajectories between the classes at the beginning of the study (P_o) and the classes at the end of the study (P_f). In this case, the intervention is also shown. From this graphic, it is possible to see how the individuals belong to one or other final class depending on the intervention assigned.

In this Figure 5.15, it is possible to observe that both interventions and control are not balanced inside each class. In general, men groups are the most balanced ones (33% VOO, 29% WOO, 37% $Ctrl$), Young women groups (31.4% VOO, 43.1% WOO and 25.5% $Ctrl$) and Women with Menopause (43% VOO, 7% WOO, 50% $Ctrl$). Comparing the subclasses, YW-UH contains the greatest proportion of WOO and WM-WMwSugars the smallest.

In addition, except for class WM-WMwSugars, individuals that are assigned to an intervention are more divided in the final classes than the individuals assigned to the control group which are more concentrated in some final class indicating that the $Ctrl$ groups are more stable. As it will be seen in Section 5.7.2, WM-WMwSugars with $Ctrl$ intervention was the group that improved their diet consuming more fish and changed the consumption of the olive oil.

Then, it is proposed to compare the dietary habits of the individuals of each trajectory to see the adherence to the intervention in Section 5.7.3. In addition, as it will be seen in Section 5.8, the comparison of other attributes like the biomarkers allows detecting the effect of the intervention in the individuals involved in the particular trajectory that is analyzed. Henceforth, the trajectory map that is analyzed is Figure 5.16 which shows only those trajectories with more than one individual.

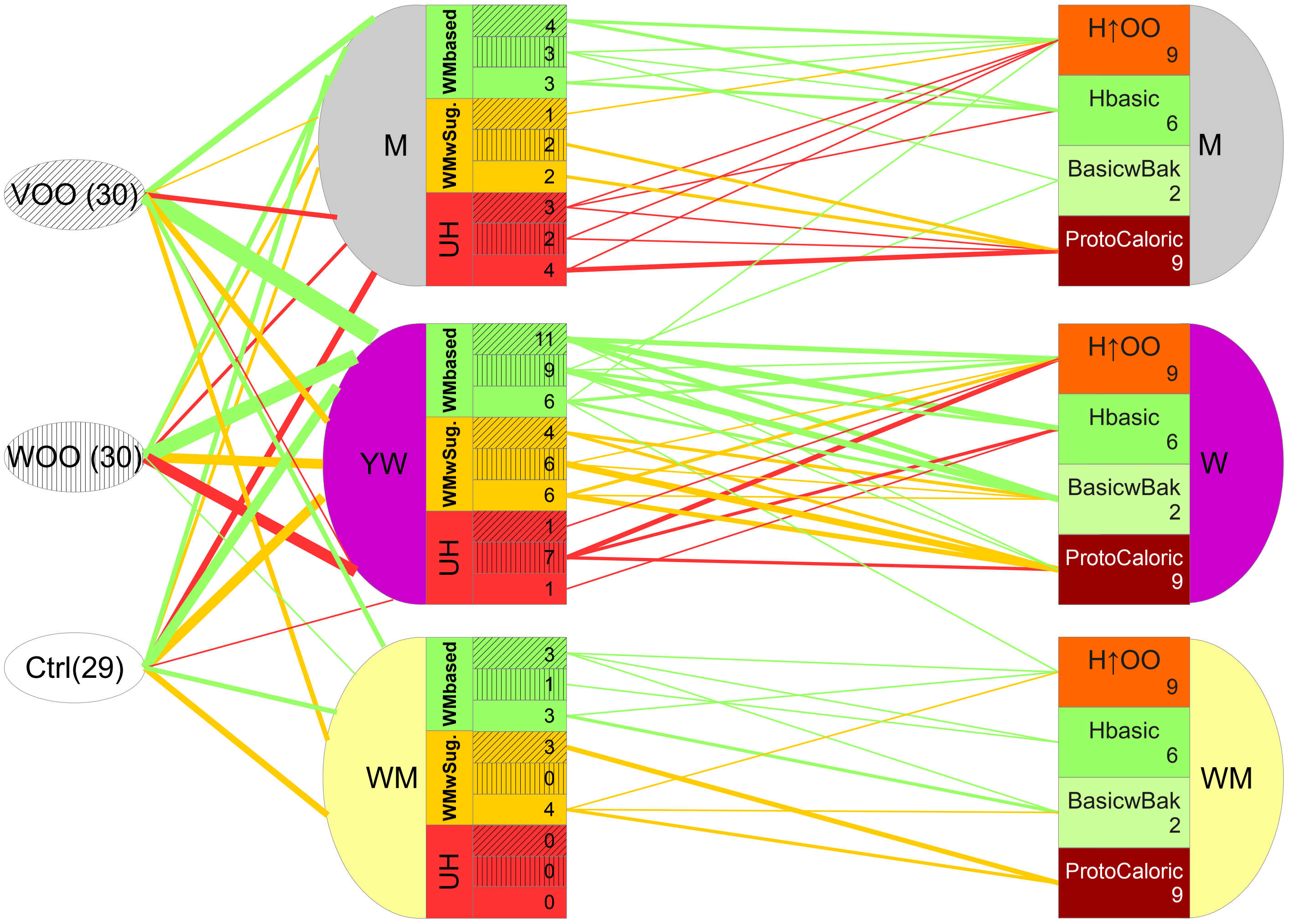


Figure 5.15: Analysis of trajectories between the PRE and POST partitions

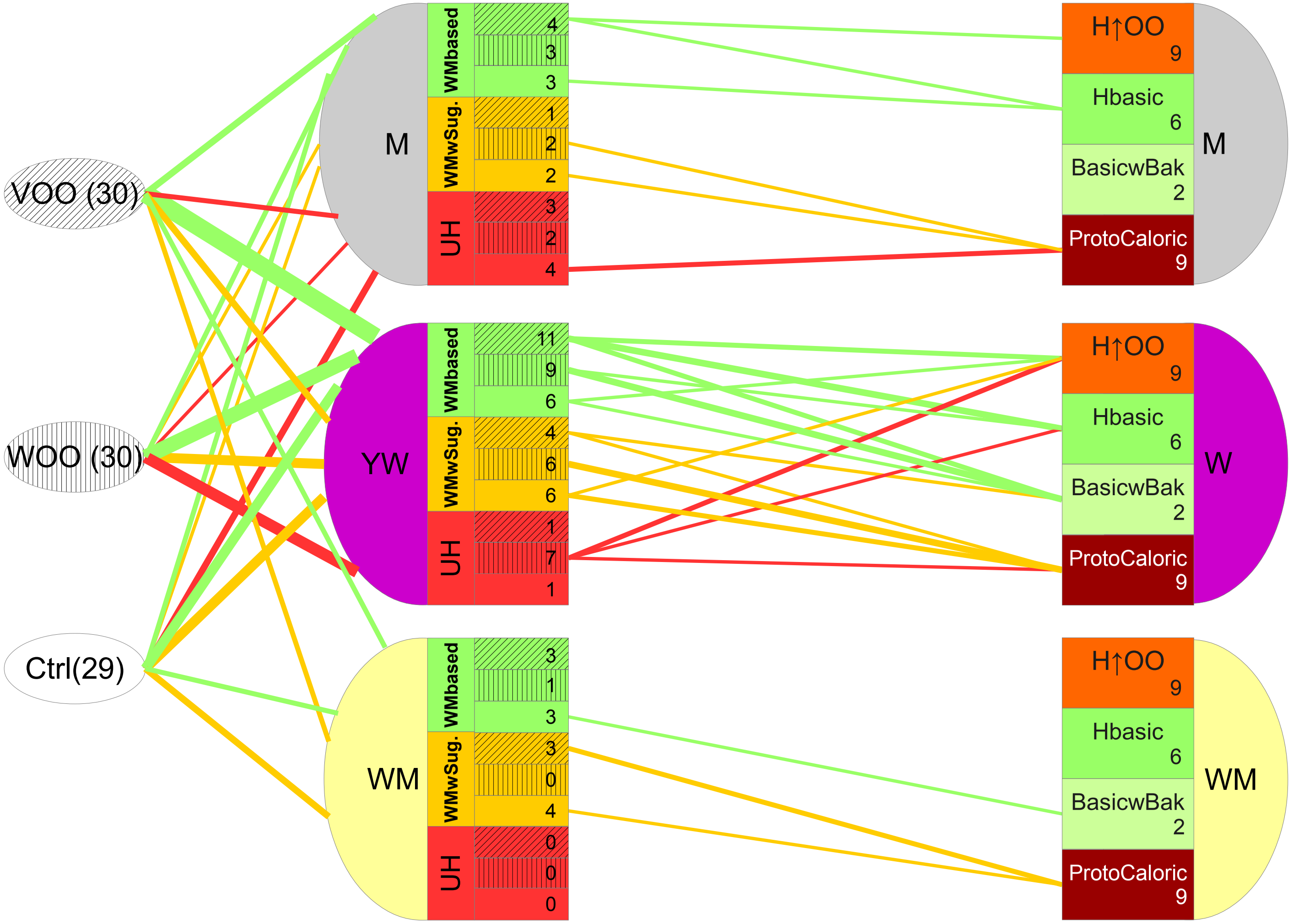


Figure 5.16: Analysis of trajectories between the PRE and POST partitions with trajectories with more than one individual

5.7 Adherence to the Intervention

In this section for evaluating the adherence to the intervention, three different approaches are presented. The first one compares the diet profiles of the different interventions. The second one, compared the diet profiles using the initial state of the individuals and the intervention ($P_o \times T$) (see Section 5.7.2) and the third one uses the found trajectories of Section 5.6 between the initial P_o and final (P_f) state of the individuals (see Section 5.7.3).

There are 3 types of intervention as explained in Section 5.1. Participants might be asked to take a Mediterranean diet with virgin olive oil (VOO); a Mediterranean diet with washed olive oil (WOO) or stay with current diet (whatever it is) in the control group (*Ctrl*). Therefore, analyzing the changes of the diet among the different groups leads to evaluate the adherence to the prescribed intervention by observing changes in diet indicators after w.r.t before the intervention.

Table 5.34: Diet Habits Attributes

Name	Alt Name	Description	Value	Semantic	Mediterranean Diet
p14_1_1	MainOliveOil	Use of olive oil as main fat	Yes/No	Maximize	Yes
p14_2_1	OliveOil	4 or more spoons/day of olive oil	Yes/No	Maximize	Yes
p14_3_1	Vegetables	2 or more pieces/day of vegetables	Yes/No	Maximize	Yes
p14_4_1	Fruit	3 or more pieces/day of fruit (including natural juices)	Yes/No	Maximize	Yes
p14_5_1	RedMeat	1 or more portions/day of red meat, hamburgers or sausages	Yes/No	Minimize	No
p14_6_1	Butter	1 or more portions/day of butter, margarine or cream	Yes/No	Minimize	No
p14_7_1	GasDrinks	1 or more glasses/day of gas drinks and/or sugary drinks	Yes/No	Minimize	No
p14_8_1	Wine	7 or more glasses/week of wine	Yes/No	Maximize	Yes
p14_9_1	Legumes	3 or more portions/week of legumes	Yes/No	Maximize	Yes
p14_10_1	Fish	3 or more portions/week of fish, shellfish or sea food	Yes/No	Maximize	Yes
p14_11_1	Commercial Bakery	2 or more portions/week of commercial bakery	Yes/No	Minimize	No
p14_12_1	Nuts	3 or more portions/week of nuts	Yes/No	Maximize	Yes
p14_13_1	WhiteMeat	Consume preferable of white meat (chicken, turkey, rabbit) instead of red meat, pork, hamburgers or sausages	Yes/No	Maximize	Yes
p14_14_1	Sauce	2 or more portions/week of homemade tomato sauce	Yes/No	Maximize	Yes

Rather than learning whether an attribute has higher or lower values; the objective is to study how these diet profiles have been changed.

Comparing the diet profiles from the beginning to the end of the intervention allows analyzing the changes of the diet patterns regarding those items that are higher or lower consumed. This information is enriched using the expert's knowledge about the meaning of the diet items. In this case, from the 14 items, 10 of them are healthy and, therefore, increasing their consumption is better, and 4 are unhealthy (better to decrease the consumption) as indicated in Table 5.34.

Therefore, the difference of two profiles can be measured by means of positive and negative effect:

- Positive effect:
  - Increased consumption of Healthy items.
  - Decreased consumption of Unhealthy items.
- Negative effect:
  - Decreased consumption of Healthy items.
  - Increased consumption of Unhealthy items.

The interpretations are made using the NCI-IMS methodology, the differences between the profiles before (instant t-1) and after the intervention (instant t) are assessed according to Table 3.13 described in Section 3.11.

5.7.1 Adherence of Intervention within each Intervention group (T)

In this section, the profiles of the diet habits before and after the study using directly intervention groups are compared in order to see the adherence to the prescribed Mediterranean diet.

Table 5.35: Dietary Difference depending on the Intervention

Intervention	p14_1_1 (mainOliveOil) - yes	p14_2_1 (oliveOil) - ≥4spoon	p14_3_1 (vegetables) - ≥2day	p14_4_1 (fruit) - ≥3day	p14_5_1 (redMeat) - ≥1day	p14_6_1 (butter) - ≥1day	p14_7_1 (gasDrinks) - ≥1day	p14_8_1 (wine) - ≥7glass/week	p14_9_1 (legume) - ≥3week	p14_10_1 (fish) - ≥3week	p14_11_1 (commercialBakery) - ≥2week	p14_12_1 (nuts) - ≥3week	p14_13_1 (whiteMeat) - yes	p14_14_1 (sauce) - ≥2week	NumIncreases	NumDecreases	Num↑Healthy	Num↓Unhealthy	size
VOO	-	-	-	-	-	-	↑	-	-	-	-	↓	-	-	1	1	0	0	30
WOO	-	-	-	-	↑	-	↑	-	↑	-	-	-	-	-	3	0	1	0	30
Ctrl	-	-	↑	-	-	-	-	-	-	↑	-	-	-	-	2	0	2	0	29

Table 5.35 shows the differences of the profiles of the individuals of each intervention group at the beginning and the end of the study. There are small changes between both pre and post profiles. Group *VOO* worsen their habits by consuming more gas drinks and less nuts. Group *WOO* improves one healthy habit (higher consumption of legumes) while worsens two (higher consumption of red meat and gas drinks); *Ctrl* group improves two healthy habits (higher consumption of vegetables and fish).

5.7.2 Adherence of Intervention within each Initial Profile (P_o)

In this section, the adherence to the prescribed intervention is assessed introducing the initial states of individuals (P_o). In this approach, this information can be used to locally analyze the interactions between nutritional profiles identified with IMC before the intervention and the type of intervention assigned to the participants. To this purpose $\mathcal{P} \times T$ is built and changes in nutritional profiles of the resulting classes in between before and after the intervention are analyzed.

Table 5.36: Differences of profiles of partition P_o depending on the Intervention

	Class	Intervention	p14_1_1 (mainOliveOil) - yes	p14_2_1 (oliveOil) - ≥4spoon	p14_3_1 (vegetables) - ≥2day	p14_4_1 (fruit) - ≥3day	p14_5_1 (redMeat) - ≥1day	p14_6_1 (butter) - ≥1day	p14_7_1 (gasDrinks) - ≥1day	p14_8_1 (wine) - ≥7glass/week	p14_9_1 (legume) - ≥3week	p14_10_1 (fish) - ≥3week	p14_11_1 (commercialBakery) - ≥2week	p14_12_1 (nuts) - ≥3week	p14_13_1 (whiteMeat) - yes	p14_14_1 (sauce) - ≥2week	NumIncreases	NumDecreases	Num↑Healthy	Num↓Unhealthy	Size
1	M-WMbased	VOO	-	-	-	-	-	-	↑	-	-	-	-	↓	-	-	1	1	0	0	4
2	M-WMbased	WOO	-	-	-	↑	-	-	↓	-	-	-	-	-	-	-	1	1	1	1	3
3	M-WMbased	Ctrl	-	-	↑	-	-	-	-	-	-	-	-	-	-	-	1	0	1	0	3
4	M-WMwSugars	WOO	-	-	-	-	↑	-	-	-	-	-	-	-	-	-	1	0	0	0	2
5	M-WMwSugars	Ctrl	-	-	-	-	-	-	-	-	-	↑	-	-	-	-	1	0	1	0	2
6	M-UH	VOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	3
7	M-UH	WOO	-	-	-	-	-	-	-	↓	-	-	-	-	-	-	0	1	0	0	2
8	M-UH	Ctrl	↑	↓	-	-	-	-	-	-	-	-	-	↓	-	-	1	2	1	0	4
9	YW-WMbased	VOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	11
10	YW-WMbased	WOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	9
11	YW-WMbased	Ctrl	-	-	-	-	-	-	-	-	-	-	↑	-	-	-	1	0	0	0	6
12	YW-WMwSugars	VOO	-	-	-	-	-	-	-	-	↓	-	-	-	-	-	0	1	0	0	4
13	YW-WMwSugars	WOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	6
14	YW-WMwSugars	Ctrl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	6
15	YW-UH	WOO	-	-	↑	-	-	-	↑	-	-	-	↓	-	-	-	2	1	1	1	7
16	WM-WMbased	VOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	3
17	WM-WMbased	Ctrl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	3
18	WM-WMwSugars	VOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	3
19	WM-WMwSugars	Ctrl	-	-	-	-	-	-	-	-	-	↑	-	-	-	-	1	0	1	0	4

Then, Table 5.36 shows the differences of the profiles for each class depending on the intervention. Table 5.36 is, in fact, a refinement of Table 5.35 from previous Section 5.7.1. Thus, the results of this approach (local to each class) are compared with the ones obtained under direct use of nutritional intervention (global). It can be seen that the nutritional changes identified in Table 5.35 for a certain intervention group, in fact, can be attributable to a certain nutritional profile before the intervention. For example, it can be seen that the increase of red meat consumption identified in top of Table 5.36 for WOO group is, in fact, due to the participants in group M-WMwSugars and all other groups do not really alter significantly their nutritional habits. In fact, making the analysis from this perspective issues some more changes that are masked Table 5.35 when some subgroups change in opposite sense like is the case of gas drinks decrease in M-WMbased for WOO diet.

Therefore, more focused changes are registered when the differences are analyzed with a finer partition. And adherence to diet might be evaluated. It is not expected that none of group treated through WOO increases red meat (like participants from M-WMwSugars do), gas drinks (YW-UH) or decrease wine (M-UH) from a Mediterranean diet perspective. Similarly, it is not expected that participants prescribed with VOO increase gas drinks (like participants from M-WMbased do) or decrease nuts (M-WMbased) and legumes (YW-WMwSugars).

In any case, Table 5.36 shows that some groups do not apparently change their diets during the trial (like M-UH, YW-WMbased, WM-WMbased, WM-WMwSugars treated with VOO or YW-WMbased, YW-WMwSugars treated with WOO).

5.7.3 Adherence to Intervention using the Pre-Post Trajectories

This approach proposes comparing the dietary habits of the individuals for each trajectory to see the adherence to the intervention. Besides, as it will be seen in Section 5.8, this approach is used for comparing other attributes like the biomarkers detecting the effect of the intervention in the individuals involved in each particular trajectory that is analyzed.

For this analysis, only trajectories with more than one individual are taken into account (see the Trajectory Map of Figure 5.16 in Section 5.6).

In order to show the difference for each trajectory, new profiles of the dietary habits (before and after the intervention) are assessed including only the individuals of a trajectory. Note that the trajectories are a subset of the crossing partition $\mathcal{P}_o \times \mathcal{P}_f \times T$. For this analysis the semantic of the attributes is the same than in previous Section 5.7.2 (see Table 5.34). Table 5.37 shows the differences between the profiles for each trajectory. In this table, also, it is possible to observe a refinement of the changes comparing with the results obtained in Table 5.35 where the difference of the diets of each intervention group were assessed. In this case, the individuals involved in the trajectory from YW-WMwSugars to W-ProtoCaloric of the WOO intervention are who have increased the consumption of legumes (see the comparison in Table 5.38) which is the unique change that was not detected in previous Section 5.7.2.

The trajectory from YW-UH to W-H↑OO shows the higher number of positive changes in Table 5.37 and it is shown apart in Table 5.39.

Table 5.37: Differences of dietary attributes between P_o and P_f according to the intervention

	$Class_o$	$Class_f$	Int	p14_1_1 (mainOliveOil) - yes	p14_2_1 (oliveOil) - ≥4spoon	p14_3_1 (vegetables) - ≥2day	p14_4_1 (fruit) - ≥3day	p14_5_1 (redMeat) - ≥1day	p14_6_1 (butter) - ≥1day	p14_7_1 (gasDrinks) - ≥1day	p14_8_1 (wine) - ≥7glass/week	p14_9_1 (legume) - ≥3week	p14_10_1 (fish) - ≥3week	p14_11_1 (commercialBakery) - ≥2week	p14_12_1 (nuts) - ≥3week	p14_13_1 (whiteMeat) - yes	p14_14_1 (sauce) - ≥2week	NumIncreases	NumDecreases	Num↑Healthy	Num↓Unhealthy	TrajSize
1	M-WMbased	M-HOO	VOO	-	-	-	-	-	-	↑	-	-	-	-	↓	-	-	1	1	0	0	2
2	M-WMbased	M-HB	VOO	-	-	-	-	-	↑	-	-	-	-	-	-	-	-	1	0	0	0	2
3	M-WMbased	M-HB	Ctrl	-	-	↑	-	-	-	-	-	-	-	-	-	-	-	1	0	1	0	2
4	M-WMwSugars	M-ProtoCaloric	WOO	-	-	-	-	↑	-	-	-	-	-	-	-	-	-	1	0	0	0	2
5	M-WMwSugars	M-ProtoCaloric	Ctrl	-	-	-	-	-	-	-	-	-	↑	-	-	-	-	1	0	1	0	2
6	M-UH	M-ProtoCaloric	Ctrl	↑	↓	-	-	-	-	-	-	-	-	-	↓	-	-	1	2	1	0	3
7	YW-WMbased	W-HOO	VOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	3
8	YW-WMbased	W-HOO	Ctrl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	2
9	YW-WMbased	W-HB	VOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	4
10	YW-WMbased	W-HB	WOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	2
11	YW-WMbased	W-BasicBak	VOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	3
12	YW-WMbased	W-BasicBak	WOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	4
13	YW-WMbased	W-BasicBak	Ctrl	-	-	-	-	-	-	-	-	-	-	↑	-	-	-	1	0	0	0	2
14	YW-WMwSugars	W-HOO	Ctrl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	2
15	YW-WMwSugars	W-BasicBak	VOO	-	-	-	-	-	-	-	-	↓	-	-	-	-	-	0	1	0	0	2
16	YW-WMwSugars	W-ProtoCaloric	VOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	2
17	YW-WMwSugars	W-ProtoCaloric	WOO	-	-	-	-	↑	-	-	-	↑	↓	-	-	-	-	2	1	1	0	4
18	YW-WMwSugars	W-ProtoCaloric	Ctrl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	3
19	YW-UH	W-HOO	WOO	-	↑	-	-	-	-	↑	-	-	↑	↓	-	-	-	3	1	2	1	3
20	YW-UH	W-HB	WOO	-	-	↑	-	-	-	-	-	-	-	-	-	-	-	1	0	1	0	2
21	YW-UH	W-ProtoCaloric	WOO	-	-	↑	-	-	-	↑	-	-	-	-	-	-	-	2	0	1	0	2
22	WM-WMbased	WM-BasicBak	Ctrl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	2
23	WM-WMwSugars	WM-ProtoCaloric	VOO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	3
24	WM-WMwSugars	WM-ProtoCaloric	Ctrl	-	-	-	-	-	-	-	-	-	↑	-	-	-	-	1	0	1	0	2

Table 5.38: Trajectory from YW-WMwSugars to W-ProtoCaloric of WOO intervention

YW-WMwSugars-W-ProtoCaloric-WOO	↑	↓
Healthy Items	Legumes	Fruit
Unhealthy Items	Red Meat	🟩

Table 5.39: Trajectory YW-UH to W-H↑OO of WOO intervention

YW-UH-W-H↑OO-WOO	↑	↓
Healthy Items	Olive Oil Fish	🟥
Unhealthy Items	Gas Drinks	Commercial Bakery

However, Table 5.37 allows easily observing whether the individuals of one trajectory have improved or not through the counters of healthy items that increase and of unhealthy items that decrease. Also, Table 5.38 and Table 5.39 let to identify the improvement of the individuals in the corresponding trajectory by the colors of the cells. The final Pre-Post Analysis tool shows this type of tables when a trajectory from the Trajectory Map is clicked as it is described in Section 4.3.

From Table 5.37, some groups assigned to *Ctrl* intervention have changed their habits. For instance, M-WMbased-M-HB, M-WMwSugars-M-ProtoCaloric and WM-WMwSugars-M-ProtoCaloric, they have improved one healthy habit. Nutritional experts outlined that it is common that control group improves their diet because the participants become more aware of its habits by being involved in a clinical trial.

Nevertheless, it is not possible to measure how this improvement is achieved because of the data type. As stated in next Section 5.7.4, the information from the survey p14 is not enough to detect all changes in diet.

5.7.4 On the Intrinsic Precision of Data

All presented approaches suffer the same problem. The information from the survey p14 is not enough to detect all changes in diet. This is, in fact, due to the intrinsic precision of the p-14 assessment scale provided nutritional as indicators in the data matrix. The attributes have only two categories indicating whether they consume more or less than a certain quantity in a certain time: for example, p14_4_1 registers whether a person consumes 3 or more pieces of fruit daily. Thus, all persons with VOO or WOO prescription that moves from no fruit at all before the study to 2 pieces cannot be identified even if they adopted a better diet than before, and they made greater effort that those moving from 2 to 3 who are identified as improving. However, despite of this problem, our approaches of studying the adherence locally to each class provides more information than when are globally (see Section 5.7.2 and Section 5.7.3).

At this time, a retroactive recovery process is followed for obtaining details of the participants about the quantity of each food. When more precise information on diet habits is available the proposed methodology might provide more precise view of the diet adherence, as well as of diet effect when adhesion.

Individual food quantities consumptions would be really welcomed to this purpose, but unfortunately, they were not available. By the moment of the submission, the only available additional information was consumption of olive oil in grams per day (either virgin (Voo) or washed (Woo)). These two attributes are important since are the difference between both interventions VOO and WOO interventions and the p-14 survey does not provide distinction between virgin or washed olive oil. Also, the total consumption of olive oil (TotalOO) is assessed by adding Voo and Woo.

Table 5.40 shows the averages of consumption of all olive oils and Table 5.41 register the changes per each intervention groups. From these two tables, it is possible to see that *VOO* group have changes their habits consuming more total olive oil (TotalOO) by increasing the consumption of Voo and decreasing Woo. Also, *Ctrl* group does not change significantly their habits. Nevertheless, group *WOO* does not change their habits because already they took more Woo than Voo (see Table 5.40).

Table 5.40: Averages of consumption of virgin olive oil (Voo) and washed olive oil (Woo) (gr/day). Measured before (pre) and after (post) the intervention.

Intervention:	Group: *VOO*			Group: *WOO*			Group: *Ctrl*		
Attribute:	Voo	Woo	TotalOO	Voo	Woo	TotalOO	Voo	Woo	TotalOO
Pre	13,17	20,48	33,64	9,11	26,21	35,32	23,42	13,94	37,36
Post	38,33	0,83	39,17	0	37,93	37,93	22,90	12,93	35,84

Table 5.41: Differences of virgin olive oil (Voo), washed olive oil (Woo) and total olive oil (TotalOO) in the intervention groups

	Intervention	Voo	Woo	TotalOO	NumIncreases	NumDecreases	Size
1	*VOO*	↑	↓	↑	2	1	30
2	*WOO*	-	-	-	0	0	30
3	*Ctrl*	-	-	-	0	0	29

As it was mentioned in previous sections, the local analysis for each initial state of the individuals provides more information than when global intervention groups are analyzed. Table 5.42) shows the averages of olive oil consumption for each initial state w.r.t the intervention ($\mathcal{P} \times T$).

From Table 5.42, first thing is that *Ctrl* group does not change olive oil consumption in general, but some profiles do, like M-WMwSugars that reduces Voo; YW-WMwSugars that reduces Woo; WM-WMbased that moves from Woo to Voo and in opposite sense WM-WMwSugars decreases Voo and increases Woo. All groups with *VOO* prescription moved from Woo consumption to Voo except WM-WMwSugars whereas all groups from *WOO* move from Voo to Woo or already they took more Woo than Voo (YW-WMwSugars and YW-UH).

Table 5.43 shows the registered changes on for each group of $\mathcal{P} \times T$ according to Voo, Woo and the sum of total olive oil (TotalOO). Working with quantities of Voo and Woo shows finer results than the ones achievable with P14_1_1 and P14_2_1 items as it can be seen in Table 5.43 by the TotalOO attribute that represents the quantity of total oil. In this table, the changes between profiles before and after the intervention are register. In addition, the changes previously detected in Table 5.42 are significant according this Table

Table 5.42: Averages of consumption of virgin olive oil (Voo) and washed olive oil (Woo) (gr/day). Measured before (pre) and after (post) the intervention for each class of P_o conditioned by the 3 interventions groups

Intervention:		Group: *VOO*			Group: *WOO*			Group: *Ctrl*		
Attribute:		Voo	Woo	TotalOO	Voo	Woo	TotalOO	Voo	Woo	TotalOO
M-WMBased	Pre	6,25	26,07	32,32	8,33	25,00	33,33	33,33	16,67	50,00
	Post	43,75	0	43,75	0	33,33	33,33	33,33	16,67	50,00
M-WMwSugars	Pre	-	-	-	37,50	0	37,50	12,50	12,50	25,00
	Post	-	-	-	0	37,50	37,50	5,00	12,50	17,50
M-UH	Pre	8,33	25,00	33,33	0	37,50	37,50	18,75	18,75	37,50
	Post	41,67	0	41,67	0	37,50	37,50	18,75	18,75	37,50
YW-WMbased	Pre	16,82	12,27	29,09	10,63	16,88	27,51	25,71	8,33	34,04
	Post	36,36	0	36,36	0	37,50	37,50	25,71	8,33	34,04
YW-WMwSugars	Pre	8,75	31,25	40,00	4,88	33,33	38,21	29,17	9,05	38,22
	Post	43,75	0	43,75	0	41,67	41,67	29,17	4,17	33,34
YW-UH	Pre	-	-	-	7,14	32,14	39,28	-	-	-
	Post	-	-	-	0	35,71	35,71	-	-	-
WM-WMBased	Pre	16,67	33,33	50,00	-	-	-	16,67	25,00	41,67
	Post	50,00	0	50,00	-	-	-	25,00	8,33	33,33
WM-WMwSugars	Pre	16,67	8,33	25,00	-	-	-	25,00	6,25	31,25
	Post	16,67	8,33	25,00	-	-	-	18,75	18,75	37,50

Table 5.43: Differences of virgin olive oil (Voo), washed olive oil (Woo) and total olive oil (TotalOO) for each group of $\mathcal{P}_o \times T$

	Class	Intervention	Voo	Woo	TotalOO	NumIncreases	NumDecreases	Size
1	M-WMbased	VOO	**↑**	**↓**	↑	2	1	4
2	M-WMbased	WOO	-	↑	-	1	0	3
3	M-WMbased	Ctrl	-	-	-	0	0	3
4	M-WMwSugars	WOO	**↓**	**↑**	-	1	1	2
5	M-WMwSugars	Ctrl	↓	-	-	0	1	2
6	M-UH	VOO	**↑**	↓	↑	2	1	3
7	M-UH	WOO	-	-	-	0	0	2
8	M-UH	Ctrl	-	-	-	0	0	4
9	YW-WMbased	VOO	↑	-	↑	2	0	11
10	YW-WMbased	WOO	↓	↑	↑	2	1	9
11	YW-WMbased	Ctrl	-	-	-	0	0	6
12	YW-WMwSugars	VOO	**↑**	**↓**	-	1	1	4
13	YW-WMwSugars	WOO	-	-	↑	1	0	6
14	YW-WMwSugars	Ctrl	-	-	-	0	0	6
15	YW-UH	WOO	-	-	-	0	1	7
16	WM-WMbased	VOO	↑	**↓**	-	1	1	3
17	WM-WMbased	Ctrl	↑	↓	↓	1	2	3
18	WM-WMwSugars	VOO	-	-	-	0	0	3
19	WM-WMwSugars	Ctrl	↓	↑	↑	2	1	4

5.43 except for class YW-WMwSugars with *VOO* intervention that has not significantly reduces the consumption of Woo.

Therefore, this proposal permits identifying irregularities in the adherence to the prescribed diet like the ones shows by WM-WMbased from *Ctrl* group or WM-WMwSugars from all intervention groups.

Finally, the differences are analyzed using the resulting found trajectories in Section

5.6. This refinement let identifying preciser changes on each group. For instance, class YW-WMbased of *WOO* has reduced Voo and increased Woo and TotalOO in Table 5.43 and, in Table 5.44, it can be seen that the reduction of Voo is due to the individuals with final state YW-HB and the increment of TotalOO to the individuals with final state YW-BasicBak.

One can also identify irregularities as in the previous approach where of class WM-WMBased and WM-WMwSugars assigned to *Ctrl* change their habits and yet WM-WMwSugars assigned to *VOO* does not.

Table 5.44: Differences of virgin olive oil (Voo), washed olive oil (Woo) and total olive oil (TotalOO) for each trajectory included in $\mathcal{P}_o \times T \times \mathcal{P}_o$

	$Class_o$	$Class_f$	Intervention	Voo	Woo	TotalOO	NumIncreases	NumDecreases	TrajSize
1	M-WMbased	M-HOO	VOO	↑	↓	-	1	1	2
2	M-WMbased	M-HB	VOO	↑	↓	↑	2	1	2
3	M-WMbased	M-HB	Ctrl	-	-	-	0	0	2
4	M-WMwSugars	M-ProtoCaloric	WOO	↓	↑	-	1	1	2
5	M-WMwSugars	M-ProtoCaloric	Ctrl	↓	-	-	0	1	2
6	M-UH	M-ProtoCaloric	Ctrl	-	-	-	0	0	3
7	YW-WMbased	W-HOO	VOO	↑	-	↑	2	0	3
8	YW-WMbased	W-HOO	Ctrl	-	-	-	0	0	2
9	YW-WMbased	W-HB	VOO	↑	↓	↑	2	1	4
10	YW-WMbased	W-HB	WOO	↓	↑	-	1	1	2
11	YW-WMbased	W-BasicBak	VOO	↑	↓	-	1	1	3
12	YW-WMbased	W-BasicBak	WOO	-	↑	↑	2	0	4
13	YW-WMbased	W-BasicBak	Ctrl	-	-	-	0	0	2
14	YW-WMwSugars	W-HOO	Ctrl	-	-	↓	0	1	2
15	YW-WMwSugars	W-BasicBak	VOO	↑	↓	-	1	1	2
16	YW-WMwSugars	W-ProtoCaloric	VOO	↑	↓	-	1	1	2
17	YW-WMwSugars	W-ProtoCaloric	WOO	-	-	-	0	0	4
18	YW-WMwSugars	W-ProtoCaloric	Ctrl	-	-	-	0	0	3
19	YW-UH	W-HOO	WOO	-	-	-	0	0	3
20	YW-UH	W-HB	WOO	↓	↑	-	1	1	2
21	YW-UH	W-ProtoCaloric	WOO	↓	↑	-	1	1	2
22	WM-WMbased	WM-BasicBak	Ctrl	↑	↓	↓	1	2	2
23	WM-WMwSugars	WM-ProtoCaloric	VOO	-	-	-	0	0	3
24	WM-WMwSugars	WM-ProtoCaloric	Ctrl	↓	↑	↑	2	1	2

Henceforth, the analysis of the intervention effects is made using the trajectories because it provides preciser results than using only the initial state of the individuals w.r.t the intervention.

5.8 Analysis of the Intervention Effects

In previous section, according to the consumption of Virgin olive oil and Washed olive oil the adherence to the assigned intervention seems to be positive. Although, some groups do not change their habits because they already took the prescribed olive oil except for WM-WMwSugars of *VOO* that do no change their habits. Also, some control groups have improved a little their habits. Nevertheless, the type of attributes offered by the p14 survey does not allow evaluating all changes that could occur in the dietary habits with precision.

In this section, the same methodology of the trajectory characterization is used, in order, to see the changes of other attributes. Indeed, these changes are the effect of the intervention along the study. For this task, since there many attributes, only blood/urine biomarkers and a selection of some genes are used to see the intervention effects. Section 5.8.1 includes the intervention effects of blood and urine biomarkers that have additive effect and Section 5.8.2 includes the multiplicative effects of the gene expression.

5.8.1 Blood and Urine Biomarkers

In Table 5.45, the differences between the blood and urine biomarkers profiles are shown using the trajectories from the partition P_o to P_f depending on the intervention (see the Trajectory Map in Section 5.6).

The attributes of Tyrosol (Tyrosol and OHTyrosol) are urine markers that are used to determine the quantity of polyphenol in the body. The virgin olive oil is rich in polyphenols. Although, the washed olive oil contains polyphenols, the concentration of virgin olive oil is three-times higher. It is expected to see that, at least, the group *VOO* has higher values than the other groups at the end of the study. However, as it can be seen in Table 5.46, none intervention group increase their levels when comparing the global profiles.

Table 5.46: Differences in Tyrosol attributes per Intervention

Intervention	tyru0 (Tyrosol)	ohtyru0 (OHTyrosol)	mohtyu0 (MOH_Tyrosol)	NumIncreases	NumDecreases	Num↑Positive	Num↓Negative	Size
VOO	-	-	-	0	0	0	0	30
WOO	-	-	-	0	0	0	0	30
Ctrl	↓	-	-	0	1	0	1	29

In Table 5.45, the 3 trajectories that increase their levels of Tyrosol are assigned to *VOO* intervention. Only one subgroup of *VOO* decreases their levels (YW-WMbased to W-BasicBak). In fact, most subgroups that decrease their levels are assigned to *WOO* intervention. Probably the reason is that the individuals that are in this intervention have replaced virgin olive oil for washed olive oil which is the one with lower levels of polyphenols.

Regarding to the lipoproteins (cholesterol, LDL, HDL, triglycerides, oxidized LDL), in men groups that the intervention does not improve their levels in general. The individuals of trajectory M-WMBased-M-HB-*VOO* have lower levels of total cholesterol but they have higher of oxidized LDL (as trajectory M-WMwSugars-M-ProtoCaloric-*WOO*). However, the trajectories in *Ctrl* seems to improve this aspect in men groups. In the young

Table 5.45: Blood and Urine Biomarkers effect on Trajectories

	$Class_o$	$Class_f$	Intervention	gluc0 (glucose)	cholest0 (cholesterol)	LDL_0 (LDL)	hdl0 (HDL)	tryg0 (triglycerides)	oxldl0 (Oxidized LDL)	isoprostanos_0 (F2α Isoprostanes)	ifn_g_0 (Interferon-γ)	mcp_1_0 (Monocyte Chemotactic Protein-1)	s_p_selectin_0 (sP-selectin)	s_cd40l_0 (sCD40 Ligand)	pcr0 (C-Reactive Protein)	oxo_gg_0 (8-Oxoguanine)	tnf_a_0 (Tumor necrosis factor-α)	t_pa_0 (Tissue Plasminogen Activator)	s_vcam_1_0 (Soluble cell adhesion molecules-1)	il_6_0 (Interleukin 6)	il_8_0 (Interleukin 8)	il_10_0 (Interleukin 10)	tyru0 (Tyrosol)	ohtyru0 (OHTyrosol)	mohtyu0 (MOH_Tyrosol)	NumIncreases	NumDecreases	TrajSize
1	M-WMbased	M-HOO	VOO	-	-	-	-	-	-	↓	-	-	-	-	↑	↓	-	-	-	↓	-	-	-	-	-	1	3	2
2	M-WMbased	M-HB	VOO	↑	↓	-	-	-	↑	-	-	-	-	**↑**	↓	-	-	↓	-	↑	-	↑	-	↓	-	5	4	2
3	M-WMbased	M-HB	Ctrl	**↓**	-	-	↑	-	-	-	-	-	↑	↑	-	**↓**	-	-	-	↑	-	**↑**	-	↓	↓	5	4	2
4	M-WMwSugars	M-ProtoCaloric	WOO	↓	-	-	-	-	↑	↓	-	-	-	↑	-	↓	-	-	-	**↓**	-	↓	↓	-	-	2	6	2
5	M-WMwSugars	M-ProtoCaloric	Ctrl	↓	-	-	-	-	-	↓	-	-	-	↑	-	↓	-	↓	-	-	-	-	-	-	-	1	4	2
6	M-UH	M-ProtoCaloric	Ctrl	-	↓	-	-	-	↓	-	-	-	↑	↓	**↑**	↓	-	↓	-	↓	-	-	-	-	↓	2	7	3
7	YW-WMbased	W-HOO	VOO	-	-	-	-	-	-	**↓**	-	-	**↓**	↓	↓	↓	-	↓	-	-	-	-	↑	-	-	1	6	3
8	YW-WMbased	W-HOO	Ctrl	-	-	-	↓	-	-	-	-	↓	-	↓	-	↓	-	-	-	↑	-	-	↓	↓	-	1	6	2
9	YW-WMbased	W-HB	VOO	-	-	-	↓	↓	↑	↓	-	-	-	↓	-	↓	-	-	-	↑	↓	-	-	**↑**	↑	4	6	4
10	YW-WMbased	W-HB	WOO	-	-	-	↓	-	↑	↓	-	↑	**↓**	↓	-	**↓**	-	-	-	↑	-	-	-	↑	-	4	5	2
11	YW-WMbased	W-BasicBak	VOO	-	-	↑	-	↑	↑	**↓**	-	-	↑	↑	-	-	-	-	-	-	-	↑	↓	↓	↓	6	4	3
12	YW-WMbased	W-BasicBak	WOO	-	↓	-	↓	↓	-	-	-	-	-	-	↑	-	-	↑	-	**↓**	-	-	-	-	-	2	4	4
13	YW-WMbased	W-BasicBak	Ctrl	-	-	↓	-	-	-	↓	-	-	-	**↓**	-	↓	-	-	-	**↓**	-	↑	-	-	-	1	5	2
14	YW-WMwSugars	W-HOO	Ctrl	-	-	-	-	-	-	-	-	↑	**↑**	**↑**	-	-	-	↑	-	-	-	-	-	-	↓	4	1	2
15	YW-WMwSugars	W-BasicBak	VOO	↑	-	-	-	-	↓	-	-	-	-	-	↓	↓	-	↓	-	↑	-	-	-	-	-	2	4	2
16	YW-WMwSugars	W-ProtoCaloric	VOO	-	-	-	-	↓	-	↑	-	↓	**↑**	↑	-	-	-	-	-	-	-	↑	**↑**	-	**↑**	6	2	2
17	YW-WMwSugars	W-ProtoCaloric	WOO	-	-	-	↑	↑	-	-	-	-	**↑**	**↑**	-	**↓**	-	-	↓	**↑**	-	↑	-	-	-	6	2	4
18	YW-WMwSugars	W-ProtoCaloric	Ctrl	↓	-	-	↓	-	-	↓	-	-	-	-	-	↓	-	↓	-	-	-	↑	-	↑	↑	3	5	3
19	YW-UH	W-HOO	WOO	-	-	-	-	-	-	-	-	-	↓	↓	↓	-	-	-	-	-	↑	-	↓	↓	↓	1	6	3
20	YW-UH	W-HB	WOO	-	↑	-	↑	-	-	-	-	↓	-	**↓**	-	**↓**	-	-	-	↓	-	**↑**	↓	-	-	3	5	2
21	YW-UH	W-ProtoCaloric	WOO	↓	↑	↑	-	↑	↑	↓	-	-	-	↑	-	-	-	-	-	-	-	↓	↓	-	-	5	4	2
22	WM-WMbased	WM-BasicBak	Ctrl	-	-	-	-	-	-	-	-	-	-	-	-	↓	-	-	-	-	-	-	-	-	-	0	1	2
23	WM-WMwSugars	WM-ProtoCaloric	VOO	↓	-	-	-	-	-	-	-	-	↓	-	-	**↓**	-	-	-	-	-	-	↑	-	-	1	3	3
24	WM-WMwSugars	WM-ProtoCaloric	Ctrl	-	↑	↑	-	-	-	-	-	-	-	-	↑	↓	-	↓	-	↓	-	**↓**	-	↓	-	3	5	2

women groups, from the trajectories assigned to VOO, two improves their levels (YW-WMwSugars to W-BasicBak and W-ProtoCaloric), two worsen (YW-WMbased to W-HB and W-BasicBak) and one has the same levels (YW-WMbased-W-H↑OO). Those assigned to WOO, in general, they have worse values (specially YW-WMbased-W-BasicBak). The trajectories assigned to $Ctrl$ are the ones that have less changes. Finally, in the women with menopause groups, only WM-WMwSugars-WM-ProtoCaloric-$Ctrl$ have worse values while the other two groups maintain the same levels.

Regarding to the attribute C-Reactive Protein whose levels rise in response to inflammation, the unique trajectories that have lower values are assigned to the VOO intervention (M-WMBased-M-HB, YW-WMBased-W-H↑OO, YW-WMwSugars-W-BasicBak, YW-UH-W-H↑OO).

In summary, from this Table 5.45 we can see that individuals of each trajectory have reacted differently in each intervention. However, the characterization of the trajectories allows us to observe these differences.

5.8.2 Analysis of Gene Expression

In this section, the changes in the gene expression are measured. The genetic tests were only made for 46 individuals. For this reason, only the trajectories that contains at least two individuals with values in gene expression are analyzed.

As it was mentioned in the problem definition (Section 3.1), the effect of these attributes are measured with a multiplicative model. For that reason, the attribute are log transform before to apply the CI-IMS methodology using the dynamic version of the generalized Test-Value.

The selection of 5 genes has been used to analyze effect of the intervention in the different classes. These 5 genes are reported to have some changes in the publication [Konstantinidou et al., 2010a]. The positive effect is to down-regulate them (inverse effect).

Table 5.47: Gene Expression Changes depending on the initial and final state and the Intervention

	$Class_o$	$Class_f$	Int	adrb2antes	arhgap15antes	ifngantes	il7rantes	polkantes	NumIncreases	NumDecreases	Num↑Positive	Num↓Negative	TrajSize
3	M-WMbased	M-HB	Ctrl	↓	-	↓	-	↑	1	2	0	2	2
5	M-WMwSugars	M-ProtoCaloric	Ctrl	-	↑	-	-	↑	2	0	0	0	2
7	YW-WMbased	W-HOO	VOO	↓	↓	**↓**	↓	↓	0	5	0	5	3
8	YW-WMbased	W-HOO	Ctrl	↑	↑	**↑**	-	↑	4	0	0	0	2
9	YW-WMbased	W-HB	VOO	↓	-	**↑**	-	-	1	1	0	1	4
11	YW-WMbased	W-BasicBak	VOO	↑	**↑**	-	-	-	2	0	0	0	3
13	YW-WMbased	W-BasicBak	Ctrl	-	↓	↑	↓	-	1	2	0	2	2
17	YW-WMwSugars	W-ProtoCaloric	WOO	**↓**	↓	↓	-	↓	0	4	0	4	4
18	YW-WMwSugars	W-ProtoCaloric	Ctrl	↑	**↑**	**↑**	↑	**↑**	5	0	0	0	3
19	YW-UH	W-HOO	WOO	↑	-	↓	↑	-	2	1	0	1	3
21	YW-UH	W-ProtoCaloric	WOO	↓	↑	↑	-	**↑**	3	1	0	1	2
23	WM-WMwSugars	WM-ProtoCaloric	VOO	-	-	-	↑	-	1	0	0	0	3

From Table 5.47 it is possible to observe that the most of down-regulated genes occur in groups of intervention *VOO* or *WOO*. In men groups, both trajectories are assigned to *Ctrl* but they have different behavior. Young women groups contains most of results. Specially, YW-WMbased-W-H↑OO-*VOO* and YW-WMwSugars-W-ProtoCaloric-*WOO* down-regulate most of the genes. The unique trajectory of women with menopause is assigned to *VOO* and have increased the gene il7r.

In Table 5.48 and Table 5.49 are shown the 45 genes. The last two columns of Table 5.49 contains the sum of all genes that have up-regulated (TotalIncreases) and all that have down-regulated (TotalDecreases). From all trajectories, YW-WMBased-W-H ↑OO-*VOO* and YW-WMwSugars-W-ProtoCaloric-*WOO* are the ones that down-regulated more genes. On the contrary, the trajectories with more up-regulated genes are YW-WMBased-W-BasicBak-*VOO*, YW-WMwSugars-W-ProtoCaloric-*Ctrl* and YW-UH-W-ProtoCaloric-*WOO*. It is interesting that YW-WMwSugars-W-ProtoCaloric have completely different behavior depending on the assigned intervention. Also, the changes in gene expression from both trajectories of YW-UH assigned to *WOO* differ depending on the diet that they follow at the end of the study.

Table 5.48: Differences of Gene Expression per Trajectory (a)

	$Class_o$	$Class_f$	Intervention	abca1antes	abcg1antes	adam17ante	adamts1antes	adrb2antes	aldh1a1antes	anxa1antes	arhgap15antes	arhgap19antes	arhgef6antes	ccng1antes	cd36antes	cetpantes	chukantes	dclre1cantes	ercc5antes	ifna1antes	ifngantes	il10antes	il6antes	il7rantes	liasantes	mpoantes	NumIncreases	NumDecreases	TrajSize
3	M-WMbased	M-HB	Ctrl	-	↓	-	↓	↓	↑	**↑**	-	-	-	-	-	-	-	↑	↓	-	↓	↑	↑	-	↓	-	5	6	2
5	M-WMwSugars	M-ProtoCaloric	Ctrl	-	-	↑	-	-	-	-	↑	-	↑	**↑**	-	-	↑	-	-	**↑**	-	**↑**	↑	-	-	-	8	0	2
7	YW-WMbased	W-HOO	VOO	-	↓	↓	↑	↓	↓	↓	↓	↓	↓	**↓**	↓	-	↓	↓	**↓**	-	**↓**	-	-	↓	↓	↓	1	17	3
8	YW-WMbased	W-HOO	Ctrl	↓	-	↑	-	↑	-	-	↑	**↑**	↑	↑	↓	-	↑	-	↑	↑	**↑**	**↑**	-	-	↑	↑	13	2	2
9	YW-WMbased	W-HB	VOO	↑	-	↑	-	↓	-	-	-	-	-	-	-	↑	-	-	-	↑	**↑**	**↑**	↑	-	-	↑	8	1	4
11	YW-WMbased	W-BasicBak	VOO	-	↑	↑	↑	↑	↑	-	**↑**	↑	**↑**	↑	↑	-	↑	↑	↑	**↑**	-	↑	↓	-	↑	-	16	1	3
13	YW-WMbased	W-BasicBak	Ctrl	**↑**	-	-	**↑**	-	↓	↑	↓	-	↓	-	-	↓	-	-	↓	-	↑	↑	↓	↓	-	-	5	7	2
17	YW-WMwSugars	W-ProtoCaloric	WOO	-	↓	↓	↓	**↓**	**↓**	↓	↓	-	-	-	↓	**↓**	↓	↓	↓	**↑**	↓	↓	↓	-	↓	↓	1	17	4
18	YW-WMwSugars	W-ProtoCaloric	Ctrl	-	↑	**↑**	**↑**	↑	-	**↑**	**↑**	**↑**	↑	**↑**	↑	↑	**↑**	**↑**	**↑**	**↑**	**↑**	↑	-	↑	↑	↑	20	0	3
19	YW-UH	W-HOO	WOO	↑	**↑**	-	-	↑	**↓**	**↓**	-	**↑**	↑	↓	↓	↓	-	↑	↑	↓	↓	↓	-	↑	↑	↓	9	9	3
21	YW-UH	W-ProtoCaloric	WOO	↑	↑	**↑**	-	↓	↑	↑	↑	↑	↑	↑	↑	↑	**↑**	↑	↑	-	↑	↑	**↑**	-	-	-	17	1	2
23	WM-WMwSugars	WM-ProtoCaloric	VOO	**↑**	-	↑	-	-	↑	↑	-	↑	↑	↑	-	-	↑	↑	-	↑	-	↑	-	↑	-	-	12	0	3

Table 5.49: Differences of Gene Expression per Trajectory (b)

	$Class_o$	$Class_f$	Intervention	msr1antes	nfkb2antes	nr1h2antes	nr1h3antes	ogtantes	olr1antes	osbpantes	pla2g4bantes	polkantes	pparaantes	pparbpantes	ppardantes	ppargantes	ptgs1antes	ptgs2antes	rgs2antes	scarb1antes	tnfsf10antes	tnfsf12_tnfsf13antes	tp53antes	usp48antes	xrcc5antes	NumIncreases	NumDecreases	TrajSize	TotalIncreases	TotalDecreases
3	M-WMbased	M-HB	Ctrl	↑	↓	-	-	↑	↓	↑	↓	↑	↓	-	↓	↑	↓	-	↑	↓	↑	-	↑	-	-	8	7	2	13	13
5	M-WMwSugars	M-ProtoCaloric	Ctrl	-	-	↓	-	-	↑	↑	↑	↑	-	-	-	-	↓	↑	↓	-	-	-	↓	↑	↑	7	4	2	15	4
7	YW-WMbased	W-HOO	VOO	↓	↓	↓	↓	↓	-	↓	↓	↓	↓	↓	-	-	-	-	↓	-	↓	-	↓	↓	↓	0	15	3	1	32
8	YW-WMbased	W-HOO	Ctrl	-	↓	↓	↑	↑	↑	↓	↑	↑	-	↑	-	-	-	-	↑	-	↓	↓	↓	↑	↑	9	6	2	22	8
9	YW-WMbased	W-HB	VOO	↓	-	↓	↓	-	↑	-	-	-	↓	-	-	-	↓	↑	↑	↓	↓	↓	-	-	-	3	8	4	11	9
11	YW-WMbased	W-BasicBak	VOO	-	↑	↑	-	↑	↓	↑	↑	-	↑	↑	↑	↑	-	↓	↑	↑	-	↑	↑	↑	↑	15	2	3	31	3
13	YW-WMbased	W-BasicBak	Ctrl	-	↓	↓	-	-	-	-	-	-	↓	-	-	-	-	-	↑	-	↑	-	-	↓	-	2	4	2	7	11
17	YW-WMwSugars	W-ProtoCaloric	WOO	-	↓	↓	-	↓	↓	↓	↓	↓	↓	↓	↓	↑	-	↓	↓	↓	↓	↓	↓	-	↓	1	17	4	2	34
18	YW-WMwSugars	W-ProtoCaloric	Ctrl	-	-	↑	↑	↑	↓	↑	↑	↑	↑	↑	↑	↑	↑	↓	↑	↑	↑	↑	↑	↑	↑	18	2	3	38	2
19	YW-UH	W-HOO	WOO	↑	↑	-	↑	↑	-	↑	↑	-	↑	↑	↑	↓	-	↓	↓	-	↓	-	-	-	↓	9	5	3	18	14
21	YW-UH	W-ProtoCaloric	WOO	↑	↑	↑	↑	-	↑	↑	↑	↑	↑	↑	↑	↑	-	↑	-	↑	↑	↑	↓	↑	↑	18	1	2	35	2
23	WM-WMwSugars	WM-ProtoCaloric	VOO	↑	-	-	-	-	↑	-	-	-	-	-	↓	↓	-	↓	↑	-	↑	-	↓	↑	↑	6	4	3	18	4

5.8.3 Description of Joint Effect Models

In this section, the compilation of all results from previous sections are presented for each trajectory.

M-WMbased-M-HOO-VOO : The individuals of this trajectory have worsened their habits by taking more gas drinks and less nuts w.r.t the Mediterranean diet. They adhere to the consumption of Voo and, also, they have decreased Woo intake.

Respect to the biomarkers, they have increased the levels of C-Reactive Protein and they have decreased their levels of F2α Isoprostanes and interleukin 6.

No gene expression is available.

M-WMbased-M-HB-VOO : They have worsened their habits by taking more butter w.r.t the Mediterranea diet. However, they adhere to the consumption of Voo (increment in Voo and TotalOO and decrement of Woo).

Respect to the biomarkers they have increased their levels of glucose, oxidized LDL, sCD40 Ligand, Interleukin 6, Interleukin 10 and they have decreased cholesterol, C-Reactive Protein, Tissue Plasminogen Activator, OH-Tyrosol.

No gene expression is available.

M-WMbased-M-HB-$Ctrl$: Although it is expected no changes, this group improves their diet introducing more vegetables.

During the study, their levels of HDL, sP-selectin, sCD40 Ligand, interleukin 6 and interleukin 10 have increased and their levels of glucose, OH-Tyrosol and MOH-Tyrosol have decreased.

Regarding the gene expression, 13 genes have up-regulated (i.e: polk) and 13 have down-regulated (i.e: adrb2, ifng).

M-WMwSugars-M-ProtoCaloric-WOO : They have worsened their habits by intaking more red meat. However, they adhere to WOO by increasing Woo and decreasing Voo consumption.

They have higher levels of oxidized LDL and sCD40 Ligand and lower levels of glucose, F2α Isoprostanes, Interleukin 6, Interleukin 10 and Tyrosol.

No gene expression is available.

M-WMwSugars-M-ProtoCaloric-$Ctrl$: They have improved their habits consuming more fish and worsened by decreasing Voo consumption (they have low consumption of olive oil in general). They have higher levels of sCD40 Ligand and lower of glucose, F2α Isoprostanes and Tissue Plasminogen Activator.

Also, they up-regulated 15 genes such as arhgap15 and polk and down-regulated 4 genes.

M-UH-M-ProtoCaloric-*Ctrl* : They consume less nuts. Their consumption of oil is the same.

Their levels of sP-selectin and C-Reactive Protein are higher and levels of cholesterol, oxidized LDL, sCD40 Ligand, Tissue Plasminogen Activator, interleukin 6 and MOH-Tyrosol are lower.

No gene expression is available.

YW-WMbased-W-HOO-*VOO* : This group already have a healthy diet and they have increased their consumption of Voo and TotalOO.

Their levels of Tyrosol are higher and their levels of F2α Isoprostanes, sP-selectin, sCD40 Ligand, C-Reactive Protein and Tissue Plasminogen Activator are lower.

They have up-regulated 1 gene (adamts1) and the have down-regulated 32 (adrb2, arhgap15, ifng, il7r, polk, etc.)

YW-WMbased-W-HOO-*Ctrl* : They have not changed their habits. However, some changes in the following biomarkers and gene expression are registered: Increase of interleukin 6 levels; decrease of levels of HDL, Monocyte Chemotactic Protein-1, Tyrosol and OH-Tyrosol. Also, 22 genes are up-regulated (adrb2, arhgap15, ifng, polk, etc.) and 8 down-regulated.

YW-WMbased-W-HB-*VOO* : Not registered changes in their diet but they have increased their consumption of Voo and TotalOO and they have decreased the consumption of Woo. They have lower levels of oxidized LDL, interleukin 6, OH-Tyrosol and MOH-Tyrosol, and higher of HDL, triglycerides, F2α Isoprostanes and sCD40 Ligand.

In addition, 11 genes are up-regulated (adrb2, ifng) and 9 down-regulated.

YW-WMbased-W-HB-*WOO* : They have adhered by decreasing Voo consumption and increasing Woo.

Their levels of oxidized LDL, Monocyte Chemotactic Protein-1, interleukin 6 and OH-Tyrosol have increased and levels of HDL, F2α Isoprostanes, sP-selectin and sCD40 Ligand have decreased.

No gene expression is available.

YW-WMbased-W-BasicBak-*VOO* : They increase Voo and decrease Woo consumption. This group shows a worse status than at the beginning of the study.

They have increased the levels of LDL, triglycerides, oxidized LDL, sP-selectin, sCD40 Ligand and interleukin 10 and they have decreased the levels of F2α Isoprostanes, Tyrosol, OH-Tyrosol and MOH-Tyrosol

Respect to the gene expression 31 genes are up-regulated (adrb2, arhgap15, etc.) and 3 down-regulated.

YW-WMbased-W-BasicBak-*WOO* : This group have increased the consumption of Woo and TotalOO. Respect to biomarkers, their levels of C-Reactive Protein and Tissue Plasminogen Activator are higher and levels of cholesterol, HDL, triglycerides and interleukin 6 are lower.

No gene expression is available.

YW-WMbased-W-BasicBak-*Ctrl* : They have worsened their habits by consuming more commercial bakery.

Their level of interleukin 10 is higher and levels of LDL, F2α Isoprostanes, sCD40 Ligand and interleukin 6 are lower.

In addition, 7 genes are up-regulated (ifng, etc.) and 11 are down-regulated (arhgap15, il7r, etc.).

YW-WMwSugars-W-HOO-*Ctrl* : Their consumption of TotalOO is lower.

Their levels of Monocyte Chemotactic Protein-1, sP-selectin, sCD40 Ligand and Tissue Plasminogen Activator are higher and levels of MOH-Tyrosol are lower.

No gene expression is available.

YW-WMwSugars-W-BasicBak-*VOO* : They have lower intake of legumes. They have increased their consumption of Voo and reduced of Woo.

Their levels of glucose and interleukin 6 are higher and their levels of oxidized LDL, C-Reactive Protein and Tissue Plasminogen Activator are lower.

No gene expression is available.

YW-WMwSugars-W-ProtoCaloric-*VOO* : They increase Voo and decrease Woo consumption.

Their levels of F2α Isoprostanes, sP-selectin, sCD40 Ligand, interleukin 10, Tyrosol and MOH-Tyrosol are higher and levels of triglycerides and Monocyte Chemotactic Protein-1 are lower.

No gene expression is available.

YW-WMwSugars-W-ProtoCaloric-*WOO* : This group worsens their habits by consuming more red meat and less fish and improves it by consuming more legumes w.r.t Mediterranean diet. They maintain a high consumption of Woo.

Their levels of HDL, triglycerides, sP-selectin, sCD40 Ligand, interleukin 6 and interleukin 10 are higher and levels of Soluble cell adhesion molecules-1 are lower.

In addition, 2 genes are up-regulated and 34 are down-regulated such as adrb2 arhgap15 ifng and polk.

YW-WMwSugars-W-ProtoCaloric-*Ctrl* : They have not changed habits.

However, their levels of interleukin 10, OH-Tyrosol and MOH-Tyrosol have increased and levels of HDL, F2α Isoprostanes and Tissue Plasminogen Activator have decreased.

In addition, 38 genes are up-regulated (adrb2, arhgap15, ifng, il7r, polk, etc.) and 2 down-regulated.

YW-UH-W-HOO-*WOO* : This group improves their habits consuming more fish and less commercial bakery and worsens by taking more gas drinks. They maintain a high consumption of Woo.

They have higher levels of interleukin 8 and lower levels of sP-selectin, sCD40 Ligand, C-Reactive Protein, Tyrosol, Oh-Tyrosol and MOH-Tyrosol.

Respect to gene expression, 18 (adrb2 il7r, etc) genes are up-regulated and 4 (ifng, etc.) are down-regulated.

YW-UH-W-HB-*WOO* : They improve by consuming more vegetables. They consume more Woo and less Voo.

They increased their levels of cholesterol, HDL and interleukin 10 and they decreased levels of Monocyte Chemotactic Protein-1, sCD40 Ligand, interleukin 6 and Tyrosol.

No gene expression is available.

YW-UH-W-ProtoCaloric-*WOO* : They consume more vegetables and more gas drinks. There is an increment in the consumption of Woo and decrement in Voo.

Their levels of cholesterol, LDL, triglycerides, oxidized LDL and sCD40 Ligand are higher and levels of glucose, F2α Isoprostanes and Tyrosol are lower.

Respect to gene expression, 35 (arhgap15, Ifng, polk, etc) genes are up-regulated and 2 (adrb2, etc.) are down-regulated.

WM-WMbased-WM-BasicBak-*Ctrl* : This group have changed their consumption of olive oil: increasing Voo and decreasing Woo and TotalOO.

However, not changes are registered in the biomarkers.

No gene expression is available.

WM-WMwSugars-WM-ProtoCaloric-*VOO* : This group have not changed any habit. However, some biomarkers registered changes: increment of Tyrosol levels and decrement of glucose and sP-selectin levels.

Respect to gene expression, 18 (il7r, etc) genes are up-regulated and 4 are down-regulated.

WM-WMwSugars-WM-ProtoCaloric-*Ctrl* : This group eats more fish. Also, they consume less Voo and more Woo and TotalOO.

Their levels of cholesterol, HDL and C-Reactive Protein have increased and levels of Tissue Plasminogen Activator, interleukin 6, interleukin 10 and OH-Tyrosol have decreased.

No gene expression is available.

5.9 Experts Assessment

The nutritional experts who provided the data have analyzed and evaluated the present case study, and their feedback is positive. They outlined that the methodology can be very valuable to analyze this type of studies considering that it provides a useful tool to find out relevant information from the data. Especially, they signaled that, even though there were not many data available, as it is usual in this kind of studies, the application of the methodology provided with insightful knowledge about the intervention study. This knowledge expressed as profiles of individuals, both at the beginning and at the end, is highly appreciated and useful for nutritional experts. They state that the profiles (for instance, the separation into man, women and women with menopause) are meaningful and match the usual medical/nutritional patterns.

Regarding the adherence to intervention issue, the experts are very interested in the analysis provided by the methodology, because these results let them to analyze how the degree of adherence to each intervention has been followed by the different profiles and sub-profiles.

The trajectory map shows, in a very understandable and visual form, a way to check how the individuals are moving from one group to another one, depending on the assigned intervention.

Regarding the intervention effects, the experts liked the information and knowledge provided by the application of the methodology, because the effects could be locally analyzed into the several profiles, and give more details than classical approaches used in the medical/nutritional field, based on global statistical tests.

5.10 Summary

In this chapter, the application of the proposed methodology over the case study (Section 5.1) has been detailed.

After the pre-processing of the data matrix (see Section 5.2), the analysis of the available attributes shows that this type of studies contains attributes with clearly different roles (see Section 5.3). The use of IMC for creating both the initial and final states has helped to obtain classes that are more easily interpretable.

The profiles of both the initial states and the final states have been important for understanding the types of individuals that are included in the study (see Section 5.4 and Section 5.5).

Analyzing the adherence to the intervention (see Section 5.7), we have seen that the data provided by the p14 survey was not sensitive enough to catch all possible changes in the prescribed diet. For that reason, a finer dataset with higher granularity is needed in order to concretely see whether the individuals are adhered better or not to the assigned intervention. At this time, a retroactive recovery process is followed for obtaining details of the quantities of each food. This new information should allow, with this same methodology, to evaluate more precisely two things: adherence to the prescribed diet (intervention) and diet effect when adhesion. However, our approaches (local to each class) gives preciser information than the one obtained under direct use of intervention groups (global). By the moment, the consumption of olive oil (either virgin or washed) is available. Using these new attribute, it has been shown that the methodology works properly and it is suitable to identify both the adherence to the intervention as possible irregularities in the adherence (see section 5.7.4).

The Trajectory Map brings a clear visualization of how the individuals behave in each initial state depending on the assigned intervention (see Section 5.6). From the characterization of the resulting trajectories, the intervention effects are easily detected. This is a general method for analyzing any type of attributes and therefore, different attributes as the biomarkers and gene expression can be easily related (see Section 5.8). Finally, the adherence and the effects of the intervention have been modeled for each trajectory.

The experts have been evaluated these results positively and state that this methodology can be very valuable to analyzed rapidly this type of studies (see more details in Section 5.9).

In next Chapter 6 are collected the experiments that have been performed on these data to validate the proposed methodology.

Chapter 6

Rationale of the Methodology

In this chapter, the experiments performed in order to evaluate the proposed Methodology for the Pre-Post intervention studies are presented.

First, the selection of the hierarchical clustering with Ward's method is evaluated comparing them with other clustering methods and using Cluster Validity indexes (see Section 6.1). Then, the performance of proposed Integrative Multiview Clustering is compared with other clustering methods. In this case, the comparative is twofold: from the structural and interpretation point of view (see Section 6.2).

Section 6.3 presents the performance of the *Test-Value* compared against classical statistical tests. Related to the Test-Value, the need of generalizing the Test-Value and the use of the Sensitive Analysisis shown in Section 6.4. In addition, the benefits of the Basic descriptors are shown in Section 6.5. In Section 6.6, the decision on how to generalize the Test-Value is addressed. Finally, the need of the consistency between the interpretation of nested partitions is shown in Section 6.7.

6.1 Selecting the Underlying Clustering Method

The goal of this section is to compare the performance of the selected hierarchical clustering based on Ward's method from a structural point of view. For this task, the obtained partition P_{H_o} (see Section 5.4.1.1) is used. Partition P_{H_o} was built using this clustering method with Gower's metrics and with the data corresponding to the 18 attributes referred to the dietary habits and physical activity at the beginning of the intervention.

In order to compare the performance of this clustering method, the following usual clustering methods have been used to create the resulting partitions to be compared with P_{H_o}:

- PAM: Partitioning of the data into k clusters "around medoids" which is a more robust version of K-means (see Section 2.3.2). A meta-algorithm has been use which tests different number of clusters. This method is built using Gower's distance and selects the best partition by the criterion of Calinski-Harabasz. The best partition is in 3 clusters (P_1 in Table 6.1).

- Hierarchical Clustering using the Complete Linkage Method which finds similar clusters [Sibson, 1973]. The criterion of selecting the two clusters ($C,\ C'$) to unify is the minimum distance between them, defining the distance as $D(C, C') = \max_{x \in C, x' \in C'} d(x, x')$. From the visualization, a convenient cut is 2 or 3 and 3 is selected (P_2 in Table 6.1).

- MClust: Model-Based Clustering. The optimal model according to BIC for EM initialized by hierarchical clustering for parameterized Gaussian mixture models [Fraley and Raftery, 2002]. The used parameters: 3 mixtures, model "EEE", ellipsoidal distribution, equal shapes and volumes, HOEM [Wilson and Martinez, 1997] distance is used. This determine that the best partition is 3 (P_3 in Table 6.1).

- DBSCAN: Generates a density based clustering of arbitrary shape [Ester et al., 1996]. In this case with an Epsilon $= 0.13$ and minimum points $= 2$ gives a partition of 6 classes (P_4 in Table 6.1)..

- Kmeans [MacQueen, 1967]. The number of clusters is computed by a meta-algorithm that determine the best partition using the HOEM distance. In this case, 3 clusters and 1000 maximum number of iterations (P_5 in Table 6.1).

Then, the resulting partitions are compared with the following indexes that evaluates the compactness and separation of the resulting clusters (see Section 2.5 for the definition of these indexes): Dunn, Normalized Hubert gamma coefficient ($\hat{\Gamma}$), Silhouettes, Calinsky-Harabasz Index (CH), Average Between-Cluster Distance, Minimum cluster separation (δ), Sindex, Average Within-Cluster distance ($\overline{W}$), C-Index and Maximum cluster diameter (Δ).

Table 6.1: Cluster Validation Indexes. The two best values are in bold.

Partition	Num.Clusters	Dunn	$\hat{\Gamma}$	Silhouettes	CH	$\overline{B}$	δ	Sindex	$\overline{W}$	C-Index	Δ
P_{H_o}	3	0.15	**0.30**	**0.13**	**13.24**	0.33	**0.09**	**0.11**	**0.27**	**0.30**	**0.58**
P_1	3	0.10	**0.30**	**0.14**	**14.72**	0.33	0.06	0.07	**0.27**	**0.29**	**0.63**
P_2	3	**0.16**	0.27	0.12	7.04	**0.36**	**0.09**	**0.11**	0.29	0.33	**0.58**
P_3	3	0.11	0.29	0.08	7.74	**0.34**	0.07	0.09	0.28	0.32	0.67
P_4	6	**0.20**	0.29	-0.08	3.23	**0.34**	**0.13**	**0.13**	0.28	0.31	0.65
P_5	3	0.04	0.10	0.03	4.25	0.32	0.03	0.05	0.30	0.44	0.71

Table 6.1 shows the 10 cluster validity indexes for each partition. The first 7 indexes indicates better quality with higher values; the last 3 with lower values.

It is possible to observe from this Table 6.1 that P_{H_o} and P_1 partitions generally perform better than the rest while P_4 and P_5 have the worst index values. For instance, partition P_4 has one cluster which contains most of the instances, that is why the diameter is so high. Also, its Dunn index (based on separation) is higher because the rest of the clusters are small and isolated. Between P_{H_o} and P_1 small differences are observed but P_{H_o} improves P_1 in 4 of the 10 indexes while P_1 improves P_{H_o} in 3.

Thus, the evaluation of the resulting clusters of Ward's method based on cluster validity indexes has been successful.

6.2 Performance of Integrative Multiview Clustering

The innovative methodology based on advanced clustering techniques has been proposed in order to find more understandable patterns or clusters (see definition in Section 3.2.1). The Integrative Multiview Clustering (IMC) combines Multiview Clustering approach with crossing operations over the several partitions obtained. In this section, a comparison with other classical clustering techniques is provided to assess the performance of our approach. The Dunn-like cluster validity index proposed by Bezdek & Pal (1998) (see description in Section 2.5) is used for the comparison from a structural point of view, as it is more robust than the original Dunn index. Also, the interpretation of the resulting partition P_o obtained in Section 5.4.2 is compared with the interpretation of other partition that has been evaluated from the structural point of view. Thus, the comparison is twofold: from the structural and interpretation point of view.

The same dataset used to build P_o is clustered under classical clustering methods for that comparison:

- Hierarchical Clustering with Ward's method. The resulting dendrogram was cut in 6 and 12 classes according to Calinski-Harabasz Index.

- PAM: a k-means approach based on medoids which is very robust (see Section 2.3.2). K was selected using a meta-algorithm that checked several values of K and proposed the one optimizing the Calinski-Harabasz Index. Here, the best cut in 8 was retained for comparison.

The Dunn-like Cluster validity index (see definition in Section 2.5) is used in order to compare all partitions (see Table 6.2).

The Integrative Multiview Clustering (IMC) obtains the best results. According to the Dunn-like Index, the clusters obtained with IMC are more compact and separable. This indicates that IMC outperforms other methods from a structural point of view.

Table 6.2: Cluster Validity Comparison. As higher value of the Dunn-like index, better is the partition.

	Num. Clusters	Dunn-like Index
Integrative MultiView Clustering (P_o)	8	0,902
Hierarchical Clustering (HC_6)	6	0,896
PAM	8	0,856
Hierarchical Clustering (HC_{12})	12	0,809

Additionally, interpretability of the clusters is also considered as it is critical for the decision-making point of view; Class Panel Graphs (CPG) and the CI-IMS interpretation methodology proposed in Section 3.6 are used to build the profiles.

In Section 5.4.5, the profiles of IMC (P_o) are detailed.

For the other obtained partitions with classical hierarchical methods and PAM, the interpretation of classes was more confusing. The best partition after P_o is HC_6. The complete interpretation of HC_6 is detailed in the Subsection 6.2.1. An Analysis of the HC_6 profiles is the following.

From the point of view of interpretation, HC_6 contains 6 classes, with some confusing structure:

- 3 of the classes (C_2, C_3, C_4) are mainly composed by women, C_5 contains basically men. C_1 shows a significantly higher proportion of men than other groups, but this seems a contradiction as the class in fact just contains 55% of men and 45% of women. Some biomarkers and food are different in C_1 and C_5. However, the conditional distributions of some of the variables pointing differences strongly overlap (Fα Isoprostanes, C-Reactive Protein, HDL, triglycerides, Tissue Plasminogen Activator) and do not support a neat characterization towards distinguishable profiles.

- On the other hand, the women classes seem to be also confusing. C_2 and C_6 seem to contain significantly older persons than C_3 and C_4. However, C_6 ranges from 20 to 70 years. The degree of overlapping of the conditional distributions is really high to assume real differences in profiles, in spite of significantly p-values obtained (glucose, Interleukin 10, Fα Isoprostanes).

- Dietary patterns show high variability in some foods that are no significant in classes.

On the one hand, IMC reduces the computational cost of analyzing high dimensional data. On the other hand, it provides more interpretable clusters. It decomposes the data into meaningful subsets that allow clustering subjects without mixing concepts, and to integrate the multiview results into a single partition that catches relationships between meaningful concepts. First, multiview principle is used to simplify the problem structure and to permit local complex models; then a further cross-clustering step integrates knowledge from all views in a single typology. Knowledge domain is used to identify the thematic blocks defining the views. The methodology is general and permits as many views as desired (which is not considered in other proposals [Bickel and Scheffer, 2004]).

From a structural point of view, the partition obtained with the Integrative Multiview Clustering is the best one according to the Dunn-like Index. Also, the interpretation of the classes, seems that the IMC obtains a clearer description of the classes when comparing with the Ward's partition (HC_6).

6.2.1 Profiles of Ward's partition HC_6

The profiles of HC_6 are the following. In the class name is indicated the size of the class.

C_1(11): A group with more proportion of Men; they are taller and heavier. Higher proportion of widows. Lower proportion of menopause and no incidence of other diseases. In general, they do not take main drugs, as global population, but the proportion of taking painkillers and "other" drugs is significantly lower. Their levels of F2α Isoprostanes, C-Reactive Protein and 8-Oxoguanine are higher. Basically diet in this group is rich in olive oil as main fat (with some using also butter), nuts, white meat, homemade sauces and poor in fruits, red meat, legumes and commercial bakery. Other foods are varying in this class.

C_2(7): Woman group that are older and smaller with higher blood pressure (systolic and diastolic) and heart rate. Higher proportion of long-term ex-smokers. Higher proportion of married and less singles. Higher proportion of menopause; higher incidence of cancer and no other diseases. Higher intake of painkillers (NSAID) and "other" drugs but not other drugs consumption. Higher levels of glucose, cholesterol, LDL, HDL, Oxidized LDL, sP-selectin, C-Reactive Protein, Tissue Plasminogen Activator and Interleukin 10. Lower levels of F2α Isoprostanes. Diet rich in olive oil as a main fat, vegetables, white meat and homemade sauces. It is poor in fruit, red meat, butter, gas drinks, wine, legumes and commercial bakery. They practice less intense exercise.

C_3(21): Women group that are smaller and thinner with lower systolic pressure. Higher proportion of smokers. Lower proportion of menopause. No incidence of other diseases. Higher intake of painkillers (NSAID) and "other drugs" but not other drugs consumption. Lower levels of glucose and Tissue Plasminogen Activator. Although they use olive oil as a main fat, they consume less olive oil than global population. The diet is rich in white meat and homemade sauce and poor in red meat, butter, gas drinks, wine, legumes and nuts. They do more home work.

C_4(15): Women group that are thinner. There is lower proportion of married and higher of divorced and separated than global population. Most have no menopause. They consider themselves more stressful and contains the unique depressed person. With no incidence of other diseases. They take more vitamins or minerals and hormones but not other drugs consumption. Their level of triglycerides is lower and level of MOH Tyrosol is higher. Their diet is rich in olive oil as a main fat, vegetables, gas drinks, wine, fish, commercial bakery, white meat and homemade sauces. It is poor in butter and legumes.

C_5(16)**:** Men group that are younger, taller and heavier. With higher systolic pressure and lower heart rate. No incidence of diseases except for dyspnea. In general, they do not take main drugs as global population but proportion of intaking painkillers (NSAID) is significant lower. Their levels of cholesterol, LDL, HDL are lower and triglycerides, Tissue Plasminogen Activator are higher. Diet rich in olive oil as a main fat, red meat, gas drinks, legumes, white meat and homemade sauces and poor in vegetables, butter and wine.

C_6(19)**:** Group mainly composed by women (79%) that are older. Higher proportion of non-smokers. Lower proportion of singles. There are more persons with menopause. No incidence of other diseases. In general they do not take main drugs as global population but proportion of intaking "other" drugs (NSAID) is significantly lower. Diet rich in olive oil as main fat, olive oil, fruit, commercial bakery, nuts, white meat and homemade sauces. Poor in vegetables, red meat, butter, gas drinks, wine and legumes.

6.3 Evaluation of the *Test-Value* Statistic

The main idea of this section is to find a connection or concordance between the interpretation that a technician/expert makes by reading *Class Panel Graphs* and a set of statistical tests expected to show significances in the same attributes and classes as experts used in the interpretation.

After a clustering process, the result is a partition that defines the different clusters or *classes* of the individuals. This resulting partition can be used as a new categorical attribute and therefore, this attribute can be compared against other attributes using statistical tests. In this work, the following tests are assessed to relate them with the interpretation of CI-IMS which is a good approximation of the obtained expert-based interpretation from direct reading of the *Class Panel Graphs*:

1. Kruskal-Wallis test (see definition is Section 2.9.1) for both active and illustrative numerical attributes.

2. χ^2-Independence test (see definition in Section 2.9.2) for both active and illustrative qualitative attributes.

3. *Test-Value* (see definition in Section 2.9.3): the initial idea was to use it only for those attributes non-significant in previous tests, but finally, it is used for both active and illustrative attributes. This test allows finding relevant attributes per each class. In addition, the behaviour of the descriptor can be derived from the statistic value. Thus, if a statistic value is negative then the mean or proportion of the class is lower than the global one, and if it is positive, it is higher.

Generally, the non-significant attributes would not be required for class description. However, in the application it will be seen how in some cases Kruskal-Wallis and χ^2-Independence test are not sensitive enough, sometimes for a high asymmetry, or too small classes or other causes. For this reason, a second test is performed: *Test-Value*, which is a new proposed strategy in cluster interpretation process.

This *Test-Value* test has been finally used for all attributes because it is a good indicator of how the classes behave in the analyzed attribute whereas χ^2-Independent test and Kruskal-Wallis test only assess the global significance of an attribute without giving precise information on which change(s) is the one having different values (that can be higher or lower than the other classes). The purpose of using this test is twofold: first, to address the lack of sensitivity that can have the previous tests for non-significant attributes and second, to add information about the behaviour of the significant attributes.

Summarizing, *Test-Value* is used to see how the attributes distinguish every single class from the general trend of the attribute. As previously indicated, this test comes from multivariate statistical techniques and we introduce it in the clustering interpretation process.

6.3.1 Experimentation

The data used in this experimentation come from our case study (see details in Section 5.1). The profiles depending on their diet and physical activity habits before the nutritional intervention are used (P_{H_o}) from Section 5.4.1.1.

The diet habits are described through 14 categorical attributes; these attributes have two categories to answer whether the person consumes more than certain quantity of certain food “per day” or “per week” depending on the type of food.

The physical activity is described through 10 numerical attributes. The level of physical activity is asked as the Kcal spent "per week" o "per year".

The resulting profiles are described in Section 5.4.4 following the Cluster Interpretation Methodolgy (CI-IMS). These profiles have been checked to be concordant according what an expert can see in the Class Panel Graphs. The corresponding Class Panel Graphs are included in Appendix B.2.2.

In the following, the performance of several statistical tests are assessed to see how they can recognize the relevant class-characteristics that are included in these profiles.

Table 6.3 shows the results of the Kruskal-Wallis test (KW) for numerical attributes and χ^2-Independence test (χ^2) for qualitative attributes against the classes. Given a risk level α, the attributes are considered significant when $p\_value < \alpha$. In this experimentation $\alpha = 0.1$ has been used. Significant attributes are marked with (*) in the table.

Table 6.3: p-values of attributes vs classes

Attribute	Type	Test	p-value	Attribute	Type	Test	p-value
mainOliveOil	Q	χ^2	0.01768*	whiteMeat	Q	χ^2	0.3712
oliveOil	Q	χ^2	0.2315	homemade sauce	Q	χ^2	0.5165
vegetables	Q	χ^2	6.572e-06*	lightWeek	N	KW	0.3625
fruit	Q	χ^2	1.97e-07*	moderateWeek	N	KW	0.02238*
redMeat	Q	χ^2	0.0001269*	intenseWeek	N	KW	0.4786
butter	Q	χ^2	0.06948*	homeWorkWeek	N	KW	0.992
gasDrinks	Q	χ^2	0.1146	totalWeek	N	KW	0.009571*
wine	Q	χ^2	0.3319	lightYear	N	KW	0.3216
legume	Q	χ^2	0.2164	moderateYear	N	KW	0.02464*
fish	Q	χ^2	0.0001266*	intenseYear	N	KW	0.2701
commercialBakery	Q	χ^2	3.248e-06*	homeWorkYear	N	KW	0.9958
nuts	Q	χ^2	0.0001091*	totalYear	N	KW	0.01987*

N: numerical Q: qualitative *: $p - value < 0.1$

These tests can identify the attributes that register important changes among classes. Significance means that the attributes have a notable different behaviour in at least one class. Table 6.4 shows the degree of concordance between attributes used by experts in their description against significance assessed by statistical tests and it can be found a good agreement. However, the drawback of these tests is that given a significant attribute, the test result itself does not provide details on which class/es is/are behaving differently than the others. For instance, the attribute *fruit* is selected as significant, but the test does not give indications on which classes contain people consuming more or less fruit. Multiple comparison techniques are required to assess these issues. This is an expensive method because a combinatorial number of tests must be performed for every significant attribute.

Table 6.4: Interpretation vs General Tests

		Kruskal-Wallis & χ^2-Indep. tests	
		Yes	No
Experts' Description	Yes	Vegetables, Fresh Fruit, Red Meat, Butter, Fish, Commercial Bakery, Nuts	gas Drinks, Legumes
	No	Main Olive Oil	Olive Oil, Wine, White Meat, Sauce

Table 6.5: Interpretation vs *Test-Value*

		Test-Value	
		Yes	No
Experts' Description	Yes	Vegetables, Fresh Fruit, Red Meat, Butter, Legumes, Gas Drinks, Fish, Commercial Bakery, Nuts	
	No	main Olive Oil	Olive Oil, Wine, White Meat, Sauce

Table 6.6: *Test-Value for both Active and Illustrative Attributes*

		Significant?		*Test-Value*			*Critical Point*		
Attribute	Type	KW/χ^2	Category	$t_{WMbased}$	t_{WMwSug}	t_{UH}	WMbased	WMwSug	UH
mainOliveOil	Q	Y	yes	1.383	0.969	-2.841*	|1.645|	|1.645|	|1.645|
oliveOil	Q	N	≥4spoon	-0.301	1.469	-1.325	|1.645|	|1.645|	|1.645|
vegetables	Q	Y	≥2day	4.777*	-2.355*	-3.22*	|1.645|	|1.645|	|1.645|
fruit	Q	Y	≥3day	-5.025*	5.084*	0.375	|1.645|	|1.645|	|1.645|
redMeat	Q	Y	≥1day	-1.424	-2.082*	4.178*	|1.645|	|1.645|	|1.645|
butter	Q	Y	≥1day	0.432	-1.997*	1.771*	|1.645|	|1.645|	|1.645|
gasDrinks	Q	N	≥1day	-0.539	-1.166	2.018*	|1.645|	|1.645|	|1.645|
wine	Q	N	≥7week	-0.368	1.328	-1.078	|1.645|	|1.645|	|1.645|
legume	Q	N	≥3week	-1.734*	0.959	1.048	|1.645|	|1.645|	|1.645|
fish	Q	Y	≥3week	2.763*	0.617	-4.15*	|1.645|	|1.645|	|1.645|
commercialBakery	Q	Y	≥2week	-4.462*	4.673*	0.149	|1.645|	|1.645|	|1.645|
nuts	Q	Y	≥3week	-0.549	3.56*	-3.433*	|1.645|	|1.645|	|1.645|
whiteMeat	Q	N	yes	-1.093	1.387	-0.243	|1.645|	|1.645|	|1.645|
homemadeSauce	Q	N	≥2week	0.383	0.568	-1.133	|1.645|	|1.645|	|1.645|
lightWeek	N	N		0.713	0.338	-1.278	|1.682|	|1.703|	|1.74|
moderateWeek	N	Y		-1.28	2.777*	-1.618	|1.682|	|1.703|	|1.74|
intenseWeek	N	N		0.792	0.28	-1.308	|1.682|	|1.703|	|1.74|
homeWorkWeek	N	N		-0.667	0.235	0.558	|1.682|	|1.703|	|1.74|
totalWeek	N	Y		0.107	2.157*	-2.627*	|1.682|	|1.703|	|1.74|
lightYear	N	N		-0.001	0.687	-0.794	|1.682|	|1.703|	|1.74|
moderateYear	N	Y		-0.721	2.353*	-1.823*	|1.682|	|1.703|	|1.74|
intenseYear	N	N		0.281	0.885	-1.373	|1.682|	|1.703|	|1.74|
homeWorkYear	N	N		-0.569	0.089	0.604	|1.682|	|1.703|	|1.74|
totalYear	N	Y		-0.197	2.201*	-2.298*	|1.682|	|1.703|	|1.74|

For each attribute, the *Test-Value* of every class is shown in Table 6.6. It can be seen that most of the attributes which are not significant for general tests (Kruskal-Wallis and χ^2-Independent test) are also not significant. Only two cases show significance with *Test-Value* which were not detected as significant by the general tests: first, the attribute *gasDrinks* is significant for class UH and secondly, *legume* is significant for the class $WMbased$.

The description of class $WMbased$ contains references to attributes: *vegetables, fresh fruit, red meat, legume and commercial bakery.* From those, all except *legume* where identified by general χ^2- Independence test, but *legume* is also retrieved when *Test-Value* is used. In fact, the *Class Panel Graph* shows a smaller proportion of people eating frequent *legumes* than the other classes and experts agree that this attribute has to be included in the description. Similarly, the Class Panel Graph show a higher proportion of frequent *gasDrinks* intake for class UH, which was not identified by general χ^2-Independence test but appears as significant with *Test-Value*.

Therefore, the results of the *Test-Value* (see Table 6.5) shows better degree of concordance between expert's criteria than the results obtained with general tests (see Table 6.4).

Table 6.7: *Summary of the Attributes Significance*

		Test-Value					*Test-Value*		
Attribute	KW/χ^2	$t_{WMbased}$	t_{WMwSug}	t_{UH}	Attribute	KW/χ^2	$t_{WMbased}$	t_{WMwSug}	t_{UH}
mainOliveOil	*		↓		whiteMeat				
oliveOil					sauce				
vegetables	*	↑	↓	↓	lightWeek				
fruit	*	↓		↑	moderateWeek	*			↑
redMeat	*		↑	↓	intenseWeek				
butter	*		↑	↓	homeWorkWeek				
gasDrinks			↑		totalWeek	*		↓	↑
wine					lightYear				
legume		↓			moderateYear	*		↓	↑
fish	*	↑	↓		intenseYear				
commercialBakery	*	↓		↑	homeWorkYear				
nuts	*		↓	↑	totalYear	*		-	↑

Table 6.7 shows a summary of the significant attributes for both global test and *Test-Value*. In column "KW/χ^2", significance is marked with "*". Significance in *Test-Value* is marked with ↑when class shows significantly higher values than the average or with ↓when it shows significantly lower values than the average. From this Table 6.7, it is possible to observe that all significant attributes for Kruskal-Wallis or χ^2-Independence tests are also significant for at least one *Test-Value*.

From the whole description provided by the proposed Test-Value, only in one case is not sensitive enough. From the *Class Panel Graph* $WMbased$ is consuming much less *redMeat* than $WMwSugars$, but the test is not detecting this fact (see CPG in Appendix B.2.2). Further analysis is required to detect these cases. See the basic descriptors in Section 6.5. The case of the attribute "Main Olive Oil" is similar. In both tests is significant because the only two individuals that have a negative answer are concentrated in class

UH. Nevertheless, a much higher proportion answer positively as the other classes. For that reason, for the general population it is characteristic that they use olive oil as a main fat. The same situation happens to the attributes *white meat* and *homemade sauces*.

Summarizing, statistical tests are introduced as a post-processing of clustering results to retrieve from clusters the relevant attributes found by the experts.

Statistical tests show that can assist or help the interpretation of the classes. As a first step classical global tests like χ^2-Independence test and Kruskal-Wallis test were used to identify relevant attributes for the clustering. However, some attributes retained as important by experts are not assessed as significant by those global tests. *Test-Value* has been imported from PCA field and used in the context of clustering, and it seems to be *more sensitive* than global tests by approaching better the experts' interpretation. They are also more expressive and give more precise information on which attribute behaves differently in which class, with a clear indication about the sense of this difference: higher or lower values than average. However, it seems that *Test-Value* is more expressive than the global tests and in no-cases the global tests are significant when the *Test-Value* is non-significant. For this reason, *Test-Value* can be directly used for interpretation, and hence, global tests can be directly skipped.

6.4 Assessment of the Sensitive Analysis of *Test-Value*

In order to analyze the effect of how the class size affects the Test-Value, the following study is performed: a sensitive analysis changing the *Test-Value* using different simulated sample sizes and maintaining the proportion of the classes respect to the sample size.

This new analysis is introduced in the interpretation procedure which enables the descriptors of a class to be characterized according to the stability of their significance to small or large variations of the sample size. The idea behind this analysis is that large differences between class means and global means will provide significant Test-Value, even with drastic reductions in sample size. In this scenario, a consistent interpretation will be obtained, whereas smaller differences might present unstable results that provide significance, or not, depending on the sample size.

From this analysis, the descriptors of the classes can be classify by its strength: more "robust" means that a reduction of class size is better withstand. The classification of the descriptors by the robustness using 6 descriptors types is proposed in Section 3.5. To create these types, two reductions and two symmetric increases of the class size are selected. Therefore, defining ϵ_1 and ϵ_2, the resulting simulated class sizes are: $\{n_c * (1 - \epsilon_1 - \epsilon_2), n_c * (1 - \epsilon_1), n_c, n_c * (1 + \epsilon_1), n_c * (1 + \epsilon_1 + \epsilon_2)\}$.

For this experimentation, the partition P_o = *{M-WMbased, M-WMwSugars, M-UH, YW-WMbased, YW-WMwSugars, YW-UH, WM-WMbased, WM-WMwSugars}* (see details of the partition in Section 5.4.2) is analyzed with this proposal using $\epsilon_1 = 0.3$ and $\epsilon_2 = 0.2$. In this analysis, a total of 64 from the 65 attributes used to build P_o are included. Attribute *work* is discarded because contains 31 categories with low probability. The *Test-Value* is assessed for each numerical attribute and, in the case of qualitative attributes:

- qualitative attribute with more than two categories: *Test-Value* for each category.
- qualitative attributes with two categories: only one category is assessed since the other is complementary and does not provide more information.

Thus, a total of 72 *Test-Value* are computed for each class. In Table 6.8 are collected the number of resulting descriptors for each type and class.

Nevertheless, when the results are analyzed, some contradictions are found. A characteristic could be significant in some class but not in other which has a higher difference with the sample mean than the difference of the first class. For example, Table 6.9 contains the local means of each class for the attribute *height* and *weight* and (*) shows those that are significant. The local mean of the *height* in M-WMwSugars is equal to the local mean of M-WMbased and thus, the difference with the global mean is the same but the first one is not significant because the size of M-WMbased is the double of M-WMwSugars size. The same occur in YW-WMwSugars and YW-UH with the *height* and in YW-WMbased

Table 6.8: Classification of descriptors per each class of P_o using the original *Test-Value*

Class	**Robust Non-descriptor ($\overline{R}$)**	**Moderate Non-descriptor ($\overline{M}$)**	**Weak Non-descriptor ($\overline{W}$)**	**Weak Descriptor (W)**	**Moderate Descriptor (M)**	**Robust Descriptor (R)**
M-WMbased	52	0	6	3	6	5
M-WMwSugars	57	4	3	3	2	3
M-UH	48	2	2	7	4	9
YW-WMbased	49	5	4	3	3	8
YW-WMwSugars	52	4	6	4	2	4
YW-UH	54	1	4	8	4	1
WM-WMbased	53	2	3	9	2	3
WM-WMwSugars	53	3	6	6	3	1
Total	418	21	34	43	26	34

Table 6.9: Means for each class of P_o where (*) mark the significance of the corresponding *Test-Value*

	Class Mean								
Attribute(X)	M-WMbased	M-WMwSugars	M-UH	YW-WMbased	YW-WMwSugars	YW-UH	WM-WMbased	WM-WMwSugars	$\overline{X}$
Height	1.73*	1.73	1.79*	1.62	1.59*	1.59	1.56*	1.58*	1.64
Weight	77.95*	78.36	90.07*	61.8*	60.3*	64.6	60.99	66.54	67.73

and WM-WMbased with *weight* where the sizes of YW-WMwSugars and YW-WMbased are greater than the size of WM-WMbased.

In order to overcome this type of inconsistencies, the *Generalized Test-Value* is introduced allowing to use the same class size for all classes (see Section 3.5). Since this option makes the *Test-Value* more sensitive, a small probability to accept significance ($\alpha = 0.01$) is used. Table 6.10 shows the number of each type of descriptors for each class of P_o.

Table 6.10: Classification of descriptors per each class of P_o using *Generalized Test-Value*

Class	Robust Non-descriptor	Moderate Non-descriptor	Weak Non-descriptor	Weak Descriptor	Moderate Descriptor	Robust Descriptor
M-WMbased	25	3	7	5	4	28
M-WMwSugars	25	6	10	5	4	22
M-UH	18	7	8	5	6	28
YW-WMbased	29	6	11	5	8	13
YW-WMwSugars	44	2	4	8	3	11
YW-UH	24	6	5	9	6	22
WM-WMbased	20	3	6	3	3	37
WM-WMwSugars	20	5	6	3	5	33
Total	205	38	57	43	39	194

6.5 Improving the Interpretation Including Basic Descriptors

For quasi-constant attributes, the introduction of a new *Test-Value* for the Poisson distribution was proposed in [Gibert et al., 2015]. Nevertheless, this option was discarded as it did not provide an improvement over the *Test-Value* for qualitative attributes.

Although analyzing all the descriptors, some apparently contradictions have been found. For example, the class M-UH contains the unique two individuals that do not use olive oil as a main fat but a 78% of this class they do. In this case, the *Test-Value* indicates that the use of olive oil as a main fat is lower in M-UH. For that reason, the introduction of the basic descriptors is proposed as it is described in Section 3.4. Then, all qualitative attributes that have a category with a higher probability than $1 - \delta$ will be included in the interpretation as Basic descriptors. If there is not a category with high probability, then the absence of the categories is assessed.

Table 6.11 shows the classification of the descriptors using the Sensitive Analysis and including the basic descriptors.

Table 6.11: Classification of descriptors per each class of P_o using CI-IMS Methodology

Class	**Robust Non-descriptor**	**Moderate Non-descriptor**	**Weak Non-descriptor**	**Weak Descriptor**	**Moderate Descriptor**	**Robust Descriptor**	**Basic Descriptor**
M-WMbased	20	5	5	3	3	15	21
M-WMwSugars	19	3	4	3	4	20	19
M-UH	17	4	3	3	3	20	22
YW-WMbased	32	1	2	7	3	6	21
YW-UH	19	2	1	7	6	14	23
WM-WMbased	17	2	4	1	4	24	20
WM-WMwSugars	17	1	4	2	2	23	23
Total	165	22	27	31	30	131	170

6.6 Improving the *Test-Value* using the Sample Size

As it is shown in Section 6.4, the original *Test-Value* produces some contradictions because it is sensible to the class size. In order to overcome these contradictions, the use of the sample size for all classes is proposed for evaluating all of them under the same conditions. The following options were explored and discarded:

- Minimum of all classes sizes: Using the minimum size is the most conservative option, but many properties are lost. $size_C = \min_{C \in P} n_c$
- Maximum of all classes sizes: $size_C = \max_{C \in P} n_c$
- Mean of the classes sizes. This option is a consensus. $\overline{size_C} = \frac{\sum_{C \in P} n_c}{\#_{C \in P}}$
- Truncated Mean: to overcome the problem of having too many small classes (satellites). The *truncated mean* is defined as a more robust option than the mean. The truncated mean is the mean of all classes that are bigger than a certain size. These small classes are defined as satellites: if $n_c < 0.05n \implies C$ is satellite. $\overline{size_P} = \frac{\sum_{C \in P, C \neq satellite} n_c}{\#_{C \in P, C \neq satellite}}$

The minimum and maximum sizes are discarded since there are many cases where the classes are unbalanced. Both mean and truncated mean depend on the number of classes, being the truncated mean more robust because do not have into account the satellite classes.

- Sample size n: this is the less conservative option. Using n the test becomes more sensitive and many characteristics appear as significant.

The option of using the sample size n guarantees that all the classes are evaluated with the same class size and also, it is independent on the number of classes or whether they are unbalanced. Since this option makes the *Test-Value* more sensitive, the probability to accept significance (α) cab be reduced. In the thesis a $\alpha = 0.01$[1] is used.

Table 6.12, Table 6.13 and Table 6.14 show the number of each descriptor-power using different methods for selecting the size of class to be used for the partitions P_{C_o}, P_{H_o} and P_o. In parentheses, the used α in the test is indicated. In those tables, it is possible to see that the original method and truncated mean have similar values. The use of the sample size n makes the test more sensitive, for that reason, the use of a more restricted probability to reject the significance can reduce this sensibility. Nevertheless, both the original and the truncated mean have a smaller mean of descriptors per class (columns 5 and 6) in Table 6.14 than in Table 6.12 and Table 6.13. That means, that the results of the truncated mean also are affected by the number of classes, since the analyzed partition P_o of Table 6.14 has 8 classes and the partitions P_{C_o} and P_{H_o} have 3 classes. In fact, the partition P_o is an specialization of others, and therefore, it is expected to have more characteristics.

Table 6.12: Count of descriptors for P_{C_o} using different methods

					Mean per Class			
Descriptor Power	**Original** ($\alpha = 0.1$)	**Truncated** ($\alpha = 0.1$)	**n** ($\alpha = 0.1$)	**n** ($\alpha = 0.01$)	**Original** ($\alpha = 0.1$)	**Truncated** ($\alpha = 0.1$)	**n** ($\alpha = 0.1$)	**n** ($\alpha = 0.1$)
Robust Non-descriptor	127	123	71	109	42.33	41.00	23.67	36.33
Moderate Non-descriptor	9	6	6	7	3.00	2.00	2.00	2.33
Weak Non-descriptor	6	7	15	14	2.00	2.33	5.00	4.67
Weak Descriptor	20	15	12	11	6.67	5.00	4.00	3.67
Moderate Descriptor	10	15	19	15	3.33	5.00	6.33	5.00
Robust Descriptor	44	50	93	60	14.67	16.67	31.00	20.00

Summarizing, the use of the sample size n in *Test-Value* is important both for comparing classes with different sizes and for not depending on the number of classes of a partition. More benefits of using the sample size are explained in Section 6.7 where the problem of nested partitions is presented.

[1] Note that until now, the used $\alpha = 0.1$

Table 6.13: Count of descriptors forP_{H_o} using different methods

					Mean per Class			
Descriptor Type	**Original** ($\alpha = 0.1$)	**Truncated** ($\alpha = 0.1$)	**n** ($\alpha = 0.1$)	**n** ($\alpha = 0.01$)	**Original** ($\alpha = 0.1$)	**Truncated** ($\alpha = 0.1$)	**n** ($\alpha = 0.1$)	**n** ($\alpha = 0.1$)
Robust Non-descriptor	155	147	106	137	51.67	49.00	35.33	45.67
Moderate Non-descriptor	7	10	6	6	2.33	3.33	2.00	2.00
Weak Non-descriptor	12	13	8	17	4	4.33	2.67	5.67
Weak Descriptor	12	14	10	13	4.00	4.67	3.33	4.33
Moderate Descriptor	13	6	17	13	4.33	2.00	5.67	4.33
Robust Descriptor	17	26	69	30	5.67	8.67	23.00	10.00

6.7 Sensitive Analysis in Nested Partitions

The objective of this section is to study the relationship between the interpretations of both nested partitions (see definition in Section 3.7). To achieve this objective, both interpretations are compared in order to tackle questions such as, given two nested partitions P^{ℓ} and $P^{\ell'}$, $P^{\ell'}$ nested in P^{ℓ}:

- Could a class be interpreted using the interpretation of its superclass?
- Could be reused the description of classes of the partition P^{ℓ} to describe the classes of $P^{\ell'}$: $Interpretation(P^{\ell'}) \subseteq Interpretation(P^{\ell'})$?
- Are the characteristics inherited from superclass of P^{ℓ} to its corresponding subclasses of $P^{\ell'}$?
- New properties arise in subclasses of partition $P^{\ell'}$?

As mentioned in Section 3.8 there are four possible situations: irrelevance, specification, inheritance and inconsistency.

Inconsistencies seem not reasonable because *Test-Value* compares the mean or proportion of the global sample against the mean/proportion of the class. Then, if a characteristic is significant for a class because this class has higher mean than the global sample. It is expected that, if the class is subdivided in subclasses, at least one subclass has significant that characteristic because exists one subclass with higher or equal mean than the class mean. In the particular case of nested partitions, it seems reasonable to expect that all

					Mean per Class			
Descriptor Type	**Original** ($\alpha = 0.1$)	**Truncated** ($\alpha = 0.1$)	**n** ($\alpha = 0.1$)	**n** ($\alpha = 0.01$)	**Original** ($\alpha = 0.1$)	**Truncated** ($\alpha = 0.1$)	**n** ($\alpha = 0.1$)	**n** ($\alpha = 0.01$)
Robust Non-descriptor	418	400	126	205	52.25	50.00	15.75	25.62
Moderate Non-descriptor	21	19	7	38	2.62	2.38	0.88	4.75
Weak Non-descriptor	34	28	29	57	4.25	3.50	3.62	7.12
Weak Descriptor	43	51	27	43	5.38	6.38	3.38	5.38
Moderate Descriptor	26	32	59	39	3.25	4.00	7.38	4.88
Robust Descriptor	34	46	328	194	4.25	5.75	41.00	24.25

Table 6.14: Count of descriptors for P_o using different methods

significant characteristics (or attributes) of a class in P_ℓ (superclass) should be inherited by the corresponding subclasses in $P_{\ell'}$ (or at least some of them). This is not always guaranteed when the original *Test-Value* (see definition in Section 2.9.3) is used in the interpretation.

An accurate analysis of which type of inconsistencies are produced by the interpretation process is described, and the reasons for the inconsistencies are analyzed (see Section 3.8). The main problem is related to the fact that when a class $C^\ell \in P_\ell$ is subdivided into several classes in $P_{\ell'}$, the sample size is reduced in the subclasses as a consequence of the subdivision; and the sensitivity of the test becomes directly affected by this fact (see Property 1 described in Section 3.3). In this way, the same distance between global and class means may be significant in the superclass, and non-significant for all subclasses, which is nonsense from the interpretative point of view.

In response to this type of paradox, the introduction of Sensitive Analysis in the interpretation procedure can help. The idea behind this analysis is that large differences between class means and global means will provide significant Test-Value, even with drastic reductions in sample size, also embracing the small sizes of the subclasses.

For exploring the relation between the interpretation of nested partitions, the initial partition P_{C_o} and the nested partition P_o are compared. The initial partition P_{C_o} of the 89 individuals is built on the basis of 47 baseline characteristics: including biometric and biochemical characteristics (see Section 5.4.1). Later, a second partition P_o is obtained as a specialization of P_{C_o} and considering 18 additional attributes of dietary and physical activity habits (see Section 5.4.2).

Partitions P_{C_o} and P_o have been interpreted by using the original Test-Value. The resulting interpretation of P_o is more detailed and some new characteristics become significant in the subclasses, but others that were relevant for the superclass are no longer significant for certain classes. The 4 situations (Irrelevance, Specification, Inheritance and Inconsistency) have been found for all attributes and Table 6.15 counts the number of descriptors in each situation.

Table 6.15: Relationships between interpretation of P_{C_o} and P_o using original *Test-Value*

	SuperClass		
SubClasses	**Non-descriptor**	**Descriptor**	Total
Non-Descriptor	106	29	135
Descriptor	36	45	81
Total	142	74	216

In Tables 6.16 and 6.17 the Class Panel Graphs of some attributes that show the different consistency patterns are shown. Table 6.16 shows the Class Panel Graphs of the attributes against partition P_{C_o}, whereas Table 6.17 against P_o. In these CPGs the significance of the original *Test-Value* is marked with ↑and ↓.

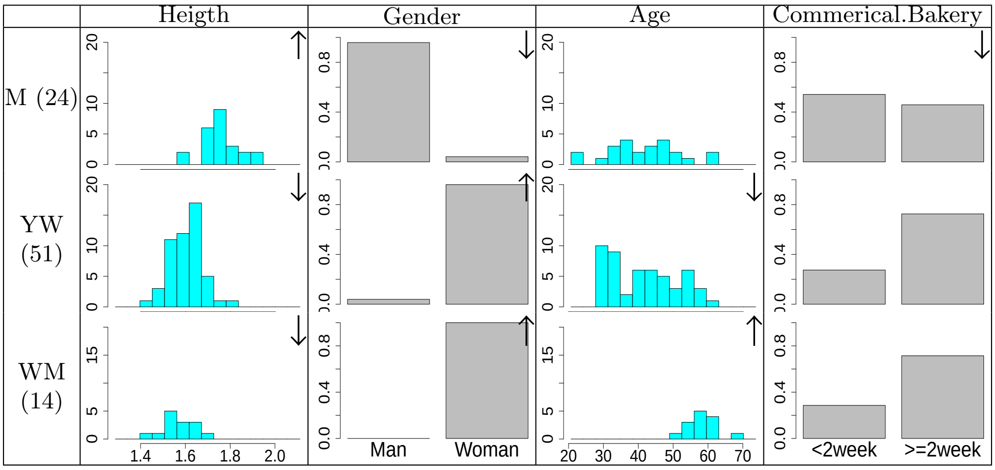


Table 6.16: Example of Class Panel Graphs

Attribute *Height* shows that some subclasses of M inherit the property of being taller than the general population, but not all, the same for classes YW and WM.

In attribute *Gender*, the inheritance for all subclasses is visible. Then, the subdivision does not contribute with additional information.

The attribute *Commercial bakery* is non-significant for class YW, but two of its subclasses are significantly higher (YW-WMwsugars and YW-UH); this is a clear case of specification. Also, in *Commercial Bakery* it is possible to observe how the class M appears significantly lower than general population, but one of the subclasses appears to

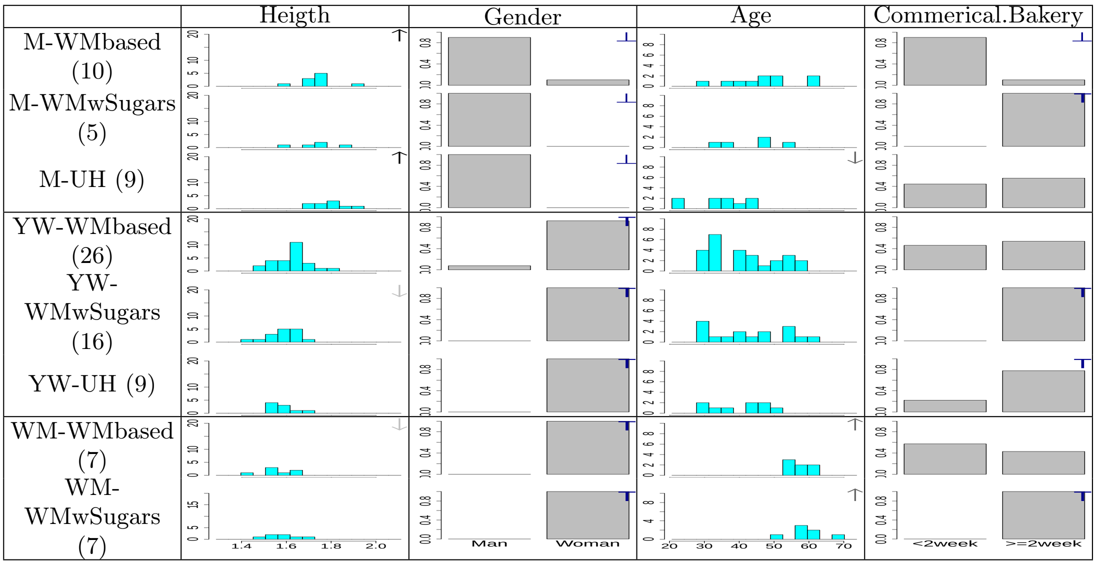


Table 6.17: Example of Class Panel Graphs

be significantly higher (M-WMwSugars). This would imply a contradiction unless some subgroups M-WMbased or M-UH appear significantly lower. This means a split of the M class in extreme distributions.

Attribute *Age* shows a case of specification of a lower *age* in M-UH class than the rest of men classes. Also, it shows that the *age* of YW appears significantly lower than general sample, but no subclass shows significant in this attribute. This is a case of inconsistency because when the group of YW shows a lower age, one expects that at least one of the subclasses of YW will concentrate all those women that are younger, and will produce a significant Test-Value. Analyzing this effect, we realize that the reduction in the sample size of subclasses reduces test sensitivity, as YW-UH shows a similar distribution to the class significant in YW.

Table 6.15, shows that there are 29 possible inconsistencies. Both partitions have been reinterpreted including the Sensitive Analysis with the original Test-Value, the relations between the interpretation are shown in Table 6.18. The results of this table are the aggregation of all situations for all subclasses of a superclass. That means that, each superclass is compared with all its subclasses at the same time.

Table 6.18: Relationships between interpretation of P_{C_o} and P_o using original interpretation methodology. (*) shows the cells that require special attention

	SuperClass						
SubClasses	**Robust Non-descriptor**	**Moderate Non-descriptor**	**Weak Non-descriptor**	**Weak Descriptor**	**Moderate Descriptor**	**Robust Descriptor**	Total
Robust Non-descriptor	85	6	2	12*	0*	0*	105
Moderate Non-descriptor	3	2	0	2*	1*	0*	8
Weak Non-descriptor	7	0	1	3*	3*	8*	24
Weak Descriptor	16	0	1	2	3	11	33
Moderate Descriptor	11	1	1	0	0	8	21
Robust Descriptor	5	0	1	1	4	16	27
Total	127	9	6	20	11	43	216

The first interesting point shown in the Table 6.18 compared with Table 3.10 in Section 3.8 is that from the 216 tests none produced hard inconsistencies between the interpretation of the superclass class and its subclasses. Moreover, most of the tests located in the remaining cells of the table correspond to inheritance, specification, or irrelevant phenomena: which provide consistent interpretations for both P_{C_o} and P_o. Only a marginal 13.4% of the tests provide situations that require attention, and of these, only a few occur in the really worrying paradoxical cells.

The most paradoxical situation is generated by the behavior of the attribute *maspirina* (aspririn consumption) for the YW class. This attribute is a moderate descriptor for class YW and a moderate or robust non-descriptor for all its subclasses.

In the following, both partitions P_{C_o} and P_o are reinterpreted using the Sensitive Analysis and the generalized version of the *Test-Value* using the sample size and a lower $\alpha = 0.01$ as it was proposed in Section 6.4. In this Section 6.4, it was discussed the selection of the simulated classes size without taking into account nested partitions. As it was mentioned, the use of the sample size was suitable because all classes are evaluated under the same size and also, it is non-dependent on the number of classes. When there is a second nested partition, in order to evaluate all the classes with the same size, the options are reduced. For instance, using the truncated mean for all classes in P_{C_o} and P_o. This solution produce dependency between both partitions because both have to be previously known to be calculated. Therefore, the use of sample size is, also, suitable for nested partitions because guarantees that all classes are evaluated under the same size and thus, the inconsistency problem is reduced and there is no need of prior knowledge of both partitions. In [Gibert et al., 2015] the results of comparing both interpretation were presented using the sample size in the *Test-Value* and with $\alpha = 0.1$. In that work, all situations that require special attention were solved but, also, most of the characteristics were descriptors at the end.

Table 6.19: Relationships between P_{C_o} and P_o using the Sensitive Analysis and the sample size in *Test-Value* with an $\alpha = 0.01$

	SuperClass						
SubClasses	**Robust Non-descriptor**	**Moderate Non-descriptor**	**Weak Non-descriptor**	**Weak Descriptor**	**Moderate Descriptor**	**Robust Descriptor**	Total
Robust Non-descriptor	28	2	0	0	0	0	30
Moderate Non-descriptor	5	2	3	0	0	0	10
Weak Non-descriptor	2	1	5	2	1	0	11
Weak Descriptor	15	0	1	1	1	0	18
Moderate Descriptor	10	0	1	0	4	0	15
Robust Descriptor	49	2	4	8	9	60	132
Total	109	7	14	11	15	60	216

Table 6.19 shows the relation of the reinterpreted classes using the Sensitive Analysis with the generalized Test-Value. This table shows a different situation to that shown in Table 6.18. It seems that almost all the inconsistencies have been completely solved,

Finally, Table 6.20 shows the results of including the Basic descriptors in the interpretation which becomes the methodology CI-IMS proposed in Section 3.6. This approach improves the concordance between both interpretations. Only 25% out of 216 test places were located in non-corner cells of the table - corresponding to the manageable situations of reinforcement and weak specification.

6.8 Summary

This chapter contains the justification of the proposed methodology.

The selection of hierarchical clustering with Ward's method have been evaluated using 10 cluster validity indexes (CVIs) in Section 6.1. This method are in the best two results of 8 of the 10 CVIs and the second best partition are in 6 CVIs.

Then, the performance of the IMC using hierarchical clustering with Ward's as an underlying method is evaluated from both the structural and interpretation points of view in Section 6.2. IMC according to the Dunn-like CVI obtains clusters more compact and separable. The interpretation of IMC seems to be clearer than with other techniques. The decomposing of the data into meaningful subsets allows clustering the individuals without mixing concepts.

Table 6.20: Relationships between P_{C_o} and P_o using the Sensitive Analysis and the sample size in *Test-Value* with an $\alpha = 0.01$

	SuperClass							
SubClasses	**Robust Non-descriptor**	**Moderate Non-descriptor**	**Weak Non-descriptor**	**Weak Descriptor**	**Moderate Descriptor**	**Robust Descriptor**	**Basic Descriptor**	Total
Robust Non-descriptor	22	1	0	0	0	0	0	23
Moderate Non-descriptor	2	0	3	0	0	0	0	5
Weak Non-descriptor	0	1	2	0	0	0	0	3
Weak Descriptor	12	0	1	0	0	0	0	13
Moderate Descriptor	9	0	1	0	1	0	0	11
Robust Descriptor	26	1	3	6	8	40	0	84
Basic Descriptor	13	1	1	2	0	5	55	77
Total	84	4	11	8	9	45	55	216

Section 6.3 compared the performance of the *Test-Value* with global statistical tests (Kruskal-Wallis test and χ^2-Independence test). The *Test-Value* is more sensitive than global statistical tests and also, it is more expressive giving more information about the attributes that behave differently, in which classes and indicating the sense of the difference.

Notwithstanding, the performance of the *Test-Value* is compromised by its sensitivity to the class size in Section 6.4. For that reason, the Sensitive Analysis is introduced to analyze the stability of a descriptor depending on small or large variations of the class size and the use of the *Generalized Test-Value* that use the same size for all the classes. In addition to the new types of descriptors, in Section 6.5, the *basic descriptors* are introduced for the categorical attributes to cope with those classes that have almost constant values. In Section 6.6, the selection of the sample size to be used in the Generalized *Test-Value* is justified because guarantees that all the classes are evaluated with the same class size and it is independent of the number of classes or whether they are unbalanced. Also, it guarantees that the same class size is used when nested partitions exists as is mentioned in Section 6.7.

In this last Section 6.7, the relationship between the interpretations of nested partitions is studied. Some inconsistencies were found when comparing both interpretation. The inconsistency problem is produced when a descriptor of a superclass disappears from all its subclasses' interpretations. To overcome these inconsistencies, a deeper analysis of the inconsistence were performed using the Sensitive Analysis, the Generalized *Test-Value* and the Basic descriptors. Most of the inconsistencies are due to the class size. Unifying these elements, the problematic inconsistencies were solved.

Next Chapter 7 includes some alternative approaches that finally have not been included in the general methodology. Specially, two alternatives for the cluster interpretation, one based in the *Test-Value* and the other based on the use of frequent itemsets.

Chapter 7

Other Research Contributions

Along this thesis some alternatives have been explored. Some of them turned in successful results and contributed to define the methodology proposed in the thesis. Others where abandoned for the final proposal. In this chapter, part of this less productive work is presented. Part of these topics are determining future lines after this PhD Thesis.

In Section 7.1, an alternative modification of the *Test-Value* is presented to detect the behavior of the attributes inside a class using its nested subclasses.

In Section 7.2, the frequent Itemsets are used for finding the profiles of the classes and a new coefficient to evaluate the itemsets is essayed. These frequent itemsets can be integrated in the Cluster Integration Methodology. Although, this experiment has been only performed for qualitative attributes, it can be a future way to explore.

Since Cluster Validity indexes are used several times, a comparison of them is performed in Section 7.3.

7.1 Local *Test-Value*

Since the Integrative Multiview Clustering gives nested partitions, each resulting class of the embedded partition has a superclass. Then, taking profit of this situation, it is proposed to use the *Test-Value* for finding differences respect to a super-class. This test compares each class (sub-class) against their super-class instead the global attribute.

The *Test-Value* is applied to all attributes comparing the class' means/proportions against both the global mean/proportion of the attribute and the local mean/proportion of its super-class (P). The *Test-Value* comparing the classes P' against its super-class P shows how some attributes behave differently inside of this super-class. Figure 7.1 shows the difference between the global and the local approach of the Test-Value.

In Table 7.1 is shown how this approach allow catch differences that the global version does not. In this table, the global version is not sensitive enough to see that the last class (WM-WMwSugars) do not take any, nevertheless using the local approach, this class is compared with its super-class WM and then, this effect is detected.

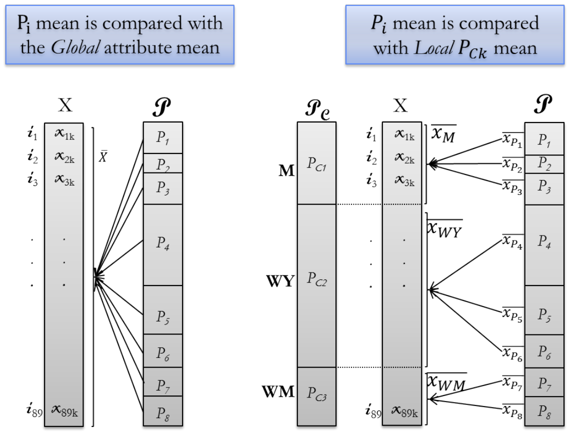


Figure 7.1: In the left part it is shown the mean of each class is compared with the mean of the whole attribute. On the right part, the mean of the class is compared with the mean of the super-class.

Table 7.1: Local *Test-Value* for the "Other" drugs qualitative attribute and the category "yes"

Proportion P_{C_o}	P_{C_o}	P_o	Proportion P_o	Global *Test-Value*	Local *Test-Value*
0.167	M	M-WMbased	0.1	-1.003	-0.741
		M-WMwSugars	0.4	0.967	-1.573
		M-UH	0.111	0.861	-0.566
0.255	YW	YW-WMbased	0.308	1.205	0.882
		YW-WMwSugars	0.125	-1.055	-1.439
		YW-UH	0.333	0.823	0.595
0.214	WM	**WM-WMbased**	0.429	1.346	**1.954***
		WM-WMwSugars	0	-1.484	**-1.954***

$\alpha = 0.1$, critical point $= 1.635$

However, this new approach let to study how a class behaves in other class. That means, in the case of Table 7.1, it is possible to see the differences of this attribute in all the classes, but also, the differences that are between the different diets and the class WM.

Summarizing, *Test-Value* is used to see how the attributes distinguish every single class from the general trend of the attribute. The *local Test-Value* is an alternative for the clustering interpretation when nested partitions occurs. Using this test both the general and local version, it is possible to see how the description of the subclasses are enriched comparing them with their corresponding superclass.

7.2 Use of Frequent ItemSets to Interpret Profiles

The description proposed for the final clustering has been built by combining the test values obtained in the classical way with an adapted *Test-Value* using local means for the P_C clustering built in Section 5.4.1 and the Class Panel Graphs. This permits in some sense to evaluate not only significant differences to the average behaviour, but also interactions of diet with the groups. Working with 8 classes and 71 attributes this process has been quite long and delicate and some criteria (conservative) have been introduced to combine the results of different tests in the final description.

In fact, one of the main interests in class interpretation is to discover local association among attributes specific for every class. That is why we have explored the possibility of using frequent itemsets to improve the interpretation process.

In this section, it is explained how frequent itemsets can be used to describe the resulting classes of a clustering.

Starting from the definition of frequent itemset in [Anand Rajaraman and Ullman, 2011], the *support* for a set of items is the proportion of subjects in the sample containing all those items. Itemsets with a support over some threshold (γ), are called *frequent itemsets*.

Then, the *Support* of an itemset $\mathcal{A}$ in a sample of size n is defined as:

$$Support(\mathcal{A}) = \frac{count\ \mathcal{A}\ in\ sample}{n} \tag{7.1}$$

In order to find frequent itemsets, all categories of attributes are examined, so as to count the number of occurrences of each possible combination of items. For typical data, with a goal of producing a small number of most frequent itemsets, the part that often spends more computational effort and size memory is the counting of each combination of items. Thus, methods for finding frequent itemsets typically concentrate on how to minimize the main memory needed to count these combinations.

One of the most popular methods to find and count frequent itemsets is the *Eclat algorithm* which is highly efficient. It is based in the following important property of itemsets (*monotonicity*): if a set of items is frequent, then so are all its subsets. This property can be exploited to eliminate the need to count certain itemsets by using its contrapositive: if an itemset is not frequent, then neither are its combination with other items.

One of the most used algorithms to find frequent itemsets is the Eclat algorithm (see Section 7.2.1). The original form of this algorithm generates all the possible combinations of items and rank the itemsets by its support. But, for our purpose: finding the frequent itemsets which describe the classes, the selection of the itemsets with higher support is not enough, because characterizing a class also requires some additional conditions. For finding the frequent itemsets to describe this class just the support of the itemset is poor, because all itemset with a high support in a class, which also has a high support in the whole sample, does not provide relevant information to describe that class. For this reason, a compromise between the frequency of an itemset within a class and within the whole sample is needed.

We propose to use the ratio of both supports inside the class and inside the total sample. This ratio shows how specific is an itemset in the class compared with the total sample. We call this ratio *Specificity Coefficient* and it will be used for finding the most representative frequent itemsets within a class. This *Specificity Coefficient* penalizes the fact that a frequent item in a class is also frequent in the whole sample.

Being $\mathcal{A}$ an itemset we can compute, besides $Support(\mathcal{A})$, the support of $\mathcal{A}$ local to a certain subset of I, namely B:

$$Support_B(\mathcal{A}) = \frac{count\ \mathcal{A}\ in\ B}{n_b} \tag{7.2}$$

$$\tag{7.3}$$

where n_B is the size of B.

The *Specificity Coefficient* of an itemset $\mathcal{A}$ given a subset B is defined as:

$$\mathcal{S}_{\mathcal{B}}(\mathcal{A}) = \frac{Support_B(\mathcal{A})}{Support(\mathcal{A})} \tag{7.4}$$

$$\tag{7.5}$$

This *Specificity Coefficient*($\mathcal{S}_{\mathcal{B}}$) is interesting in both higher and lower values. High values of $\mathcal{S}_{\mathcal{B}}(\mathcal{A})$ imply $Support_B(\mathcal{A}) > Support(\mathcal{A})$ meaning that elements satisfying $\mathcal{A}$ are mostly concentrated inside subset B. Low values of $\mathcal{S}_{\mathcal{B}}$ imply $Support_B(\mathcal{A}) < Support(\mathcal{A})$ meaning a lack of elements satisfying $\mathcal{A}$ inside B and preferentially distributed out of B.

The *Specificity Coefficient*($\mathcal{S}$) can be computed for the frequent itemsets of every class and can be used as a ranking criteria.

The Eclat algorithm is used to find frequent itemsets for all classes with γ as a *support* threshold. In addition, the maximum length of an itemset will be restricted to ℓ so as to minimize the use of main memory.

For finding longer itemsets with less computational effort, we first identify constant itemsets in every class and search frequent itemsets with the remaining attributes. This can be done on basis of the following property: Given $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_m$ itemsets such that $Support_C(\mathcal{A}_1) = Support_C(\mathcal{A}_2) = ... = Support_C(\mathcal{A}_m) = 1$ and given an itemset $\mathcal{B}$ with $Support_C(\mathcal{B}) < 1$, then: $Support_C(\mathcal{A}_1, \mathcal{A}_2, ..\mathcal{A}_m \wedge \mathcal{B}) = Support_C(\mathcal{B})$.

For this reason the itemset with $Support_C \geq \gamma$ and length $\leq \ell$ are found for C among the attributes not constant in the class.

Then, taking into account the described considerations, we propose an *algorithm in order to find these representative frequent itemsets of a class.* For each class C:

1. Set γ, ℓ for class C
2. Select the data corresponding to class C: I_C
3. Find itemsets with $Support_C = 1$ within I_C: Build $\mathcal{A}_\mathcal{C}$ as the conjunction of all the found itemsets.
4. Remove the attributes involved in $\mathcal{A}_\mathcal{C}$ from I_C: I'_C
5. Apply Eclat algorithm to I'_C restricting the support with γ and length of itemsets with ℓ and obtain a new set of itemsets: $\mathcal{B}_\mathcal{C} = \{b_{1c}, b_{2c}, ..., b_{zc}\}$
6. Build the set of frequent itemsets as $\mathcal{B}'_\mathcal{C} = \{\mathcal{A}_\mathcal{C} \wedge b_{1c}, \mathcal{A}_\mathcal{C} \wedge b_{2c}, ..., \mathcal{A}_\mathcal{C} \wedge b_{zc}\} \cup \mathcal{A}_\mathcal{C}$
7. Rank $\mathcal{B}'_\mathcal{C}$ by the *Specificity Coefficient*$(\mathcal{S}_\mathcal{C})$

From the set of itemsets $(\mathcal{B}'_\mathcal{C})$, those with higher *Specificity Coefficient* are selected to be candidates in the description of the class. For example, a *Specificity Coefficient*≥ 1.2 means that the support within the class is at least 20% superior than the total support in the sample.

As it was mentioned before, a very small value of this coefficient can mean that the absence of this itemset explains the class, but this will be a future research work.

7.2.1 Eclat Algorithm

Eclat is one of the best known basic algorithms for mining frequent itemsets in a set of transactions (subjects).

Studies of Frequent Itemset (or pattern) Mining is acknowledged in the data mining field because of its broad applications in mining association rules, correlations, and graph pattern constraint based on frequent patterns, sequential patterns, and many other data mining tasks. Efficient algorithms for mining frequent itemsets are crucial for mining association rules as well as for many other data mining tasks [Agrawal et al., 1993].

First of all, the data is transformed into a binary matrix where each column correspond to one category of one attribute, therefore the size of the matrix is the total number of categories of all attributes per the size of the sample (n).

The Eclat implementation proposed by Borgelt [Borgelt, 2003] uses (sparse) bit matrices to represent transactions lists and to filter closed and maximal item sets. To structure the search, the algorithm organizes the subset of items as a prefix tree. In this tree those item sets are combined in a node which have the same prefix with regard to some arbitrary, but fixed order of the items. Then, the search of the frequent itemsets are a depth-first order. That is, it extends an item set prefix until it reaches the boundary between frequent and infrequent item sets and then backtracks to work on the next prefix.

7.2.2 Case Study

In order to test the performance of the proposal, two classes (YW-WMbased and WM-WMwSugars) from the clustering P_o (see Section 5.4.2) are analyzed regarding the diet attributes.

The description provided in Section 5.4.5 is compared with the results of the proposed algorithm for both classes. The algorithm is applied with $\ell = 1$ and $\ell = 4$ and γ=0.8.

7.2.2.1 Class YW-WMbased

In class YW-WMbased, from the profile description of Section 5.4.5, their diet is richer in vegetables (*p14_3_1*) and fish (*p14_10_1*) and poorer in fruit (*p14_4_1*) and commercial Bakery (*p14_11_1*).

Table 7.2 shows the frequent itemsets resulting from the Eclat algorithm using the data of the class "YW-WMbased" (I_C) with $\gamma = 0.8$ and $\ell = 1$. The itemsets of this table are ordered by the $Support_C$.

Table 7.2: Frequent Itemsets of the class YW-WMbased

Number of Items	**ItemSet** (A_C)	$Support_C$
1	p14_1_1 (mainOliveOil)=yes	1
1	p14_13_1 (whiteMeat)=yes	0.9615
1	p14_14_1 (sauce)=≥2week	0.9231
1	p14_3_1 (vegetables)=≥2day	0.8462
1	p14_6_1 (butter)=<1	0.8462
1	p14_8_1 (wine)=<7glass/week	0.8462
1	p14_9_1 (legume)=<3week	0.8462
1	p14_5_1 (redMeat)=<1	0.8077
1	p14_10_1 (fish)=≥3week	0.8077
1	p14_4_1 (fruit)=<3day	0.8077

Following the explained algorithm, next Table 7.3 shows the frequent itemsets for a $\gamma = 0.8$ and $\ell = 1$. The resulting itemsets are ordered by the *Specificity Coefficient*. Remaining that the resulting itemsets are combined with the itemsets with $Support_C = 1$.

Table 7.3: Representative Frequent Itemsets for class YW-WMbased with $\gamma = 0.8$ and $\ell = 1$ including the items with $Support_C = 1$

Class	**NumItems**	**ItemSet** (I_{C_i})	$Support_C$	$Support_T$	**Spec.Coef.**
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_3_1 (vegetables)=≥2day	0.8462	0.5056	1.6735
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_4_1 (fruit)=<3day	0.8077	0.5281	1.5295
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_10_1 (fish)=≥3week	0.8077	0.5955	1.3563
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_8_1 (wine)=<7glass/week	0.8462	0.7753	1.0914

Table 7.3: Representative Frequent Itemsets for class YW-WMbased with $\gamma = 0.8$ and $\ell = 1$ including the items with $Support_C = 1$

Class	**NumItems**	**ItemSet** (I_{C_i})	$Support_C$	$Support_T$	**Spec.Coef.**
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_9_1 (legume)=<3week	0.8462	0.7865	1.0758
YW-WMbased	1	p14_1_1 (mainOliveOil)=yes	1	0.9775	1.023
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_13_1 (whiteMeat)=yes	0.9615	0.9438	1.0188
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1	0.8077	0.7978	1.0125
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_6_1 (butter)=<1	0.8462	0.8427	1.0041
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_14_1 (sauce)=≥2week	0.9231	0.9326	0.9898

The first three items with the highest *Specificity Coefficient* that are combined with "mainOliveOil (*p14_1_1*)= yes" are: more vegetables (*p14_3_1*), less fruit (*p14_4_1*) and more fish (*p14_10_1*). These three items matches the previous description.

In addition, this itemset *p14_1_1* (mainOliveOil)=yes has a *Specificity Coefficient* close to 1 because also is a frequent itemset of the whole sample and then, despite it describes either the particular class and the whole sample, it is not so relevant for the class profile.

Table 7.4 shows the itemsets after applying the proposed algorithm but with $\ell = 4$. In that case, the results are similar. The second and third itemsets contain the first itemset but combined with attributes that have a high support either inside the class and inside the sample. In addition, these two itemsets do not improve neither the support nor the *Specificity Coefficient*.

Table 7.4: Representatice Frequent Itemsets for class YW-WMbased with $\gamma = 0.8$ and $\ell = 4$ including the items with $Support_C = 1$

Class	**NumItems**	**ItemSet** (I_{C_i})	$Support_C$	$Support_T$	**Spec.Coef.**
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_3_1 (vegetables)=≥2day	0.8462	0.5056	1.6735
YW-WMbased	3	p14_1_1 (mainOliveOil)=yes p14_3_1 (vegetables)=≥2day p14_13_1 (whiteMeat)=yes	0.8077	0.4831	1.6717
YW-WMbased	3	p14_1_1 (mainOliveOil)=yes p14_3_1 (vegetables)=≥2day p14_14_1 (sauce)=≥2week	0.8077	0.4944	1.6337
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_4_1 (fruit)=<3day	0.8077	0.5281	1.5295
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_10_1 (fish)=≥3week	0.8077	0.5955	1.3563
YW-WMbased	3	p14_1_1 (mainOliveOil)=yes p14_8_1 (wine)=<7glass/week p14_13_1 (whiteMeat)=yes	0.8462	0.764	1.1075

Table 7.4: Representatice Frequent Itemsets for class YW-WMbased with $\gamma = 0.8$ and $\ell = 4$ including the items with $Support_C = 1$

Class	NumItems	ItemSet (I_{C_i})	$Support_C$	$Support_T$	Spec.Coef.
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_8_1 (wine)=<7glass/week	0.8462	0.7753	1.0914
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_9_1 (legume)=<3week	0.8462	0.7865	1.0758
YW-WMbased	3	p14_1_1 (mainOliveOil)=yes p14_9_1 (legume)=<3week p14_14_1 (sauce)=≥2week	0.8077	0.764	1.0571
YW-WMbased	3	p14_1_1 (mainOliveOil)=yes p14_9_1 (legume)=<3week p14_13_1 (whiteMeat)=yes	0.8077	0.764	1.0571
YW-WMbased	3	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1 p14_13_1 (whiteMeat)=yes	0.8077	0.7865	1.0269
YW-WMbased	1	p14_1_1 (mainOliveOil)=yes	1	0.9775	1.023
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_13_1 (whiteMeat)=yes	0.9615	0.9438	1.0188
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1	0.8077	0.7978	1.0125
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_6_1 (butter)=<1	0.8462	0.8427	1.0041
YW-WMbased	3	p14_1_1 (mainOliveOil)=yes p14_6_1 (butter)=<1 p14_14_1 (sauce)=≥2week	0.8077	0.809	0.9984
YW-WMbased	3	p14_1_1 (mainOliveOil)=yes p14_6_1 (butter)=<1 p14_13_1 (whiteMeat)=yes	0.8077	0.809	0.9984
YW-WMbased	2	p14_1_1 (mainOliveOil)=yes p14_14_1 (sauce)=≥2week	0.9231	0.9326	0.9898
YW-WMbased	3	p14_1_1 (mainOliveOil)=yes p14_13_1 (whiteMeat)=yes p14_14_1 (sauce)=≥2week	0.8846	0.8989	0.9841

7.2.2.2 Class WM-WwSugars

Class WM-WwSugars is a group that has a diet poor in vegetables (*p14_3_1*) and rich in fruit (*p14_4_1*), commercialBakery (*p14_11_1*) and nuts (*p14_12_1*) as it is described in Section 5.4.5.

In Table 7.5, the resulting itemsets from the Eclat algorithm are shown.

Table 7.5: Frequent Itemsets of the class WM-WwSugars

Number of Items	ItemSet (A_C)	$Support_C$
1	p14_1_1 (mainOliveOil)=yes	1
1	p14_13_1 (whiteMeat)=yes	1
1	p14_5_1 (redMeat)=<1	1
1	p14_6_1 (butter)=<1	1
1	p14_7_1 (gasDrinks)=<1	1

Table 7.5: Frequent Itemsets of the class WM-WwSugars

Number of Items	**ItemSet** (A_C)	*Support*$_C$
1	p14_11_1 (commercialBakery)=≥2week	1
1	p14_14_1 (sauce)=≥2week	0.8571
1	p14_12_1 (nuts)=≥3week	0.8571
1	p14_8_1 (wine)=<7glass/week	0.8571
1	p14_4_1 (fruit)=≥3day	0.8571
1	p14_3_1 (vegetables)=<2day	0.8571

First, the proposed algorithm is applied with $\gamma = 0.8$ and $\ell = 1$. Table 7.6 shows the resulting itemsets ordered by the *Specificity Coefficient*. Notice that in this class 6 attributes are constant and will be part combined with the rest of frequent itemsets.

Table 7.6: Representative Frequent Itemsets for class WM-WwSugars $\gamma = 0.8$ and $\ell = 1$ combined with the 6 itemsets with *Support*$_C = 1$

Class	**NumItems**	**ItemSet** (I_{C_i})	*Support*$_C$	*Support*$_T$	**Spec.Coef.**
WM-WwSugars	7	p14_1_1 (mainOliveOil)=yes p14_3_1 (vegetables)=<2day p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_11_1 (comlBakery)=≥2week p14_13_1 (whiteMeat)=yes	0.8571	0.1685	5.0857
WM-WwSugars	7	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_11_1 (comlBakery)=≥2week p14_12_1 (nuts)=≥3week p14_13_1 (whiteMeat)=yes	0.8571	0.1685	5.0857
WM-WwSugars	7	p14_1_1 (mainOliveOil)=yes p14_4_1 (fruit)=≥3day p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_11_1 (comBakery)=≥2week p14_13_1 (whiteMeat)=yes	0.8571	0.2472	3.4675
WM-WwSugars	7	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_8_1 (wine)=<7glass/week p14_11_1 (comlBakery)=≥2week p14_13_1 (whiteMeat)=yes	0.8571	0.2921	2.9341

Table 7.6: Representative Frequent Itemsets for class WM-WwSugars $\gamma = 0.8$ and $\ell = 1$ combined with the 6 itemsets with $Support_C = 1$

Class	**NumItems**	**ItemSet** (I_{C_i})	$Support_C$	$Support_T$	**Spec.Coef.**
WM-WwSugars	6	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_11_1 (comlBakery)=≥2week p14_13_1 (whiteMeat)=yes	1	0.3483	2.871
WM-WwSugars	7	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_11_1 (comlBakery)=≥2week p14_13_1 (whiteMeat)=yes p14_14_1 (sauce)=≥2week	0.8571	0.3258	2.6305

In this case, the first three itemsets match with the description given in Section 5.4.5. In this table, the *Specificity coefficient's* are higher than in class YW-WMbased. This is because it is easier to have a high support within a small class and a smaller support in the total sample than when the size of the class is greater.

The results of applying the algorithm increasing the itemset length to $\ell = 4$ are shown in Table 7.7 including the items with $Support_C = 1$.

The first itemset with 8 items is interesting because one of its items legume (*p14_9_1*) has a higher *specificity coefficient* combined. Nevertheless, the support of this item in most of the groups is similar and therefore, although it is true that describes the class, this attribute does not provide new information that is known of the overall sample.

Table 7.7: Representative Frequent Itemsets for class WM-WwSugars $\gamma = 0.8$ and $\ell = 4$ combined with the 6 itemsets with $Support_C = 1$

Class	**NumItems**	**ItemSet** (I_{C_i})	$Support_C$	$Support_T$	**Spec.Coef.**
WM-WwSugars	8	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_8_1 (wine)=<7glass/week p14_11_1 (comBakery)=≥2week p14_12_1 (nuts)=≥3week p14_13_1 (whiteMeat)=yes	0.8571	0.1573	5.449
WM-WwSugars	7	p14_1_1 (mainOliveOil)=yes p14_3_1 (vegetables)=<2day p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_11_1 (comlBakery)=≥2week p14_13_1 (whiteMeat)=yes	0.8571	0.1685	5.0857

Table 7.7: Representative Frequent Itemsets for class WM-WwSugars $\gamma = 0.8$ and $\ell = 4$ combined with the 6 itemsets with $Support_C = 1$

Class	NumItems	ItemSet (I_{C_i})	$Support_C$	$Support_T$	Spec.Coef.
WM-WwSugars	7	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_11_1 (comlBakery)=≥2week p14_12_1 (nuts)=≥3week p14_13_1 (whiteMeat)=yes	0.8571	0.1685	5.0857
WM-WwSugars	8	p14_1_1 (mainOliveOil)=yes p14_4_1 (fruit)=≥3day p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_11_1 (comlBakery)=≥2week p14_13_1 (whiteMeat)=yes p14_14_1 (sauce)=≥2week	0.8571	0.236	3.6327
WM-WwSugars	7	p14_1_1 (mainOliveOil)=yes p14_4_1 (fruit)=≥3day p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_11_1 (comlBakery)=≥2week p14_13_1 (whiteMeat)=yes	0.8571	0.2472	3.4675
WM-WwSugars	7	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_8_1 (wine)=<7glass/week p14_11_1 (comlBakery)=≥2week p14_13_1 (whiteMeat)=yes	0.8571	0.2921	2.9341
WM-WwSugars	6	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_11_1 (comlBakery)=≥2week p14_13_1 (whiteMeat)=yes	1	0.3483	2.871
WM-WwSugars	7	p14_1_1 (mainOliveOil)=yes p14_5_1 (redMeat)=<1 p14_6_1 (butter)=<1 p14_7_1 (gasDrinks)=<1 p14_11_1 (comlBakery)=≥2week p14_13_1 (whiteMeat)=yes p14_14_1 (sauce)=≥2week	0.8571	0.3258	2.6305

Along these case studies, we can see that frequent itemsets with very high support in both the whole sample and the class are not representative enough to describe the class.

These frequent itemsets are not meaningful for describing a particular class because those do not provide new information about the specific class. That is the reason to use another index as the *specificity coefficient* because it contributes to find the most representative itemsets for a class.

As further research line to explore is comparing the supports inside the class with the local support for P_{C_o} (resulting partition described in Section 5.4.1).

Besides, other research line to explore is searching for the less frequent itemsets in a class. An itemset with a very low *Specificity Coefficient* could be interesting for describing the class because is described by absence.

7.3 Cluster Validity Indexes

Cluster validation in Clustering is an open problem. One validation possibility is to use cluster validity indexes (CVIs). However, there are many indexes available, and they perform inconsistently scoring different partitions over a given dataset. The aim of the study carried out is the analysis of seventeen CVIs to get a common understanding of its nature, and proposing an efficient strategy for validating a given clustering. A deep understanding of what CVIs are measuring has been achieved by rewriting all of them under a common notation (see Section 2.5). This exercise revealed that indexes measure different structural properties of the clusters. As it will be seen, a Principal Component Analysis (PCA) confirmed this conceptual classification. In this study is proposed to perform a multivariate joint analysis of the indexes to learn about the cluster topology instead of using them for simple competitive ranking.

The different clustering approaches and the different configurations including the selection of the number of clusters when required by the algorithm, lead to different solutions for the same dataset [Duda et al., 2012]. Therefore, to evaluate which partition is correct or better than others becomes a crucial task.

Usually, the real partition of the data is unknown and therefore the results from a clustering process cannot be compared with a reference partition by computing misclassification indexes, as in the case of supervised learning.

In the literature, most of the techniques used for evaluating the clustering results are based on numerical indexes which evaluate the validity of the resulting partition from different points of view, known as Cluster Validity Indexes (CVI).

A wide number of cluster validity indexes can be found in literature and it is possible to find some reviews of those indexes which compare them [Arbelaitz et al., 2013, Brun et al., 2007, Dimitriadou et al., 2002, Maulik and Bandyopadhyay, 2002, Milligan and Cooper, 1985, Dubes, 1987, Halkidi et al., 2001, Halkidi et al., 2002a, Halkidi et al., 2002b]. However, there are currently no clear guidelines for deciding which is the most suitable index for a given dataset [Brun et al., 2007, Dimitriadou et al., 2002, Maulik and Bandyopadhyay, 2002, Milligan and Cooper, 1985]. In fact, there is not an agreement among those indexes, but it seems clear that each one can give some information about a different property of the partition like homogeneity, compactness of classes, variability, etc. All these cluster validity Indexes refer to structural properties of the partition, which are context-independent, and the evaluation based on them is mainly made in terms of topology.

In this section, some of these indexes are analyzed with some additional indicators of the structure of the partitions provided by statistical packages (*fpc* [Hennig, 2013a]). A new methodology is proposed based on the joint multivariate interpretation of all these indexes that provides valuable information about the topology of the classes and constitutes a richer method of evaluation.

This work differs from those found in literature because it does not perform a simple comparison among several indexes to see which is the best index. Instead, a multivariate analysis of indexes is performed to better understand the nature of those indexes and which of them have similar behavior. Since, in most of clustering real applications, one has no idea about the structure of the best partition which fits the data, we think that, at least, the proposed approach brings a valuable information about the properties of the clustering recognized.

The works found in literature (see Section 2.5) run a sort of competition among indexes to search for a winner. However, our belief is that most of these indexes are measuring different properties of the clusters, and that all of them are related with structural characteristics of the classes. It is very probable that for certain structures some indexes perform better than others. The main limitation here is that in most of clustering real applications, one has no idea about sphericity of classes or whether are tangent classes nor other features that could help to select the best index. Instead, we think that an interesting reverse reading of this scenario is suitable, by using all those indexes to get knowledge about classes' structure based on their joint performance, and this is one of the contributions of this work: the idea of making a joint interpretation of all indexes to get structural knowledge from the partition.

Concretely, a principal component analysis has been performed in order to analyze the relationships among different indexes and in this way, to establish the methodology for further analysis of a real dataset.

7.3.1 Methodology

The 17 most commonly used CVIs and indicators (see Section 2.5) found in the literature are evaluated over 17 UCI datasets previously classified. The multivariate relationships among indexes are analyzed by means of a Principal Component Analysis (PCA) (see Section 2.8). The first 2D and 3D factorial subspaces are analyzed to identify groups of indexes behaving similarly over several datasets and conclusions are extracted about how to use those indexes to get information about classes' topology.

Finally, these principles are used to learn about the structure of a real dataset from a nutritional domain. As it is well known that clustering results depend on the algorithm, data is clustered using different clustering methods. The obtained partitions are evaluated using the joint-interpretation of the index set and used to select the best partition. For the particular application presented here three clustering methods are used:

- Ward's method (see definition in Section 2.3.1)
- PAM (see definition in Section 2.3.2)
- Integrative Multiview Clustering (IMC) (see description in Section 3.2.1)

In addition, all clustering methods are distance-based methods where comparisons between objects are used to guide clusters' formation. Since we are facing a dataset that contains heterogeneous attributes, Gower's dissimilarity coefficient is used (see Section 2.4).

7.3.2 Results

A contribution of this work is that, after rewriting all the indexes under a common notation, we can observe that indexes evaluate a reduced set of characteristics of a partition, as described in Table 7.8. Thus, all indexes can be grouped around 4 basic concepts:

- Indexes measuring compactness of clusters
- Indexes measuring separation between clusters
- Indexes measuring relationships between compactness and separation
- Indexes measuring chaos in the clusters

Table 7.8: Classification of the 17 index according to their target

Index	Target	Range	Optimal Value
Entropy	Chaos in the classes	[0,)	Minimum
Diameter (Δ)	Compactness	[0,)	Minimum
WG	Compactness	[0,)	Minimum
$\overline{W}$	Compactness	[0,)	Minimum
WSS	Compactness	[0,)	Minimum
$\overline{B}$	Separation	[0,)	Maximum
Separation (δ)	Separation	[0,)	Maximum
Sindex	Separation	[0,)	Maximum
Dunn	Separation vs Compactness	[0,)	Maximum
Dunn-like	Separation vs Compactness	[0,)	Maximum
CH	Separation vs Compactness	[0,)	Maximum
$\hat{\Gamma}$	Separation vs Compactness	$[-1, 1]$	Maximum
Silhouettes	Compactness vs Separation to the nearest cluster	$[-1, 1]$	Maximum
BH	Compactness & Separation	$[-1, 1]$	Maximum
WBR	Compactness vs Separation	[0, 9	Minimum
C-Index	Compactness & Separation	$[0, 1]$	Minimum
DB	Compactness vs Separation	[0,)	Minimum

For this analysis, 17 datasets from the UCI Machine Learning Repository have been used. See in Table 7.9 the main characteristics of these datasets.

The mentioned 17 CVIs and indicators have been computed for every dataset using the real partition of each one (see Table 7.8).

Table 7.9: 17 dataset from the UCI repository and the main characteristics

Dataset	Num. Clusters	Data Type	Num. Attributes	Num. Instances
breast_w	2	Numerical	30	569
Ionosphere	2	Numerical	34	351
Parkinsons	2	Numerical	22	195
sonar_all	2	Numerical	60	208
Transfusion	2	Numerical	4	748
Haberman	2	Numerical	3	306
Musk	2	Numerical	166	476
Spectf	2	Numerical	44	267
Iris	3	Numerical	4	150
Wine	3	Numerical	13	178
vertebral_column	3	Numerical	6	310
Vehicle	4	Numerical	18	846
breast_tissue	6	Numerical	9	106
Glass	6	Numerical	9	214
Ecoli	8	Numerical	7	336
vowel_context	11	Numerical	10	990
movement_libras	15	Numerical	90	360

A Principal Component Analysis (PCA) has been performed over Table 7.10 in order to understand the relationships among different indexes. The eigenvalues recommend to keep 3 factors with a total conserved inertia of 80.2 % (see Figure 7.2).

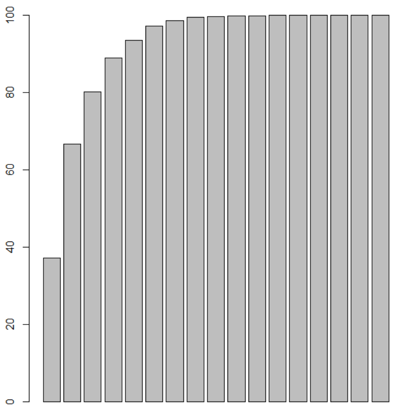


Figure 7.2: Histogram of the Inertia of the Principal Components

Figure 7.3a shows the first factorial plane. Figure 7.3b shows the 3-D projection over first three factorial axes. The indexes are represented as projected vectors over projection

Table 7.10: The 17 indexes computed for the 17 UCI datasets

		Chaos		Compactness			Separation			Relational								
Dataset	Num Clusters	Entropy	Δ	wg	$\overline{W}$	WSS	$\overline{B}$	δ	Sindex	Dunn	Dunn-like	CH	$\hat{\Gamma}$	Silhouette	BH	WBR	C-Index	DB
breast_w	2	0,66	4350,55	1145,67	382,43	$1,2110^{8}$	1030,1	10,922	18,52	0,003	1,32	633,63	0,49	0,51	0,64	0,37	0,17	0,91
ionosphere	2	0,65	9,75	5,29	3,58	3086,17	4,47	0,467	0,79	0,048	0,89	17,75	0,27	0,15	0,29	0,80	0,33	4,42
parkinsons	2	0,56	487,14	202,93	108,82	$2,1810^{6}$	140,68	3,269	4,66	0,007	0,97	13,05	0,16	0,19	0,26	0,77	0,43	3,49
sonar_all	2	0,69	3,53	1,53	1,77	351,59	1,83	0,507	0,59	0,144	1,02	6,00	0,06	0,03	0,06	0,97	0,46	5,85
transfusion	2	0,55	12250,48	3500,13	1188,38	$1,5110^{9}$	1558,94	0,000	0,00	0,000	0,86	37,42	0,11	0,18	0,12	0,76	0,47	4,26
Haberman	2	0,58	64,03	23,69	15,93	$5,3110^{4}$	17,18	0,000	0,60	0,000	0,98	8,36	0,06	0,06	0,08	0,93	0,48	5,43
Musk	2	0,68	2604,64	1437,21	1429,7	$5,2110^{8}$	1433,84	288,376	403,54	0,111	0,95	7,37	0,01	0,01	-0,00	1,00	0,49	7,87
Spectf	2	0,51	230,33	142,11	87,24	$1,110^{6}$	76,46	22,956	26,41	0,100	0,85	14,82	-0,1	-0,07	-0,19	1,14	0,67	2,77
Iris	3	1,10	3,82	0,91	0,96	89,3	3,32	0,224	0,35	0,058	1,57	487,33	0,68	0,50	0,88	0,29	0,05	0,84
Wine	3	1,09	1000,03	133,22	192,22	$5,2310^{6}$	434,53	4,785	6,87	0,005	0,73	206,68	0,42	0,20	0,52	0,44	0,18	1,87
vertebral_column	3	1,03	427,19	298,80	46,62	$4,7410^{5}$	64,54	3,020	5,90	0,007	0,56	97,71	0,21	0,11	0,38	0,72	0,29	2,36
Vehicle	4	1,39	747,47	97,01	186,52	2.4410^{7}	228,83	7,071	12,43	0,009	0,87	72,72	0,11	-0,09	0,18	0,81	0,34	14,0
breast_tissue	6	1,78	173088,92	134397,35	8694,6	$2,7210^{10}$	11426	12,630	27,56	0,000	0,02	6,73	0,04	-0,19	0,28	0,76	0,19	4,78
Glass	6	1,51	10,15	5,94	2,05	911,2	3,2	0,157	0,21	0,015	0,29	19,70	0,25	-0,09	0,38	0,64	0,27	4,26
Ecoli	8	1,52	1,07	0,84	0,32	21,31	0,63	0,052	0,07	0,049	0,45	81,18	0,61	0,24	0,80	0,50	0,11	1,70
vowel_context	11	2,40	7113,94	2214,04	2412,79	$3,3410^{9}$	3165,12	427,639	651,67	0,060	0,67	54,20	0,22	0,01	0,44	0,76	0,25	4,51
movement_libras	15	2,71	4,26	2,63	1,72	627,23	2,33	0,169	0,49	0,040	0,60	15,21	0,24	0,02	0,46	0,74	0,29	4,06
Target		Min	min	min	min	min	max	max	max	max	max	max	1	1	1	min	min	min

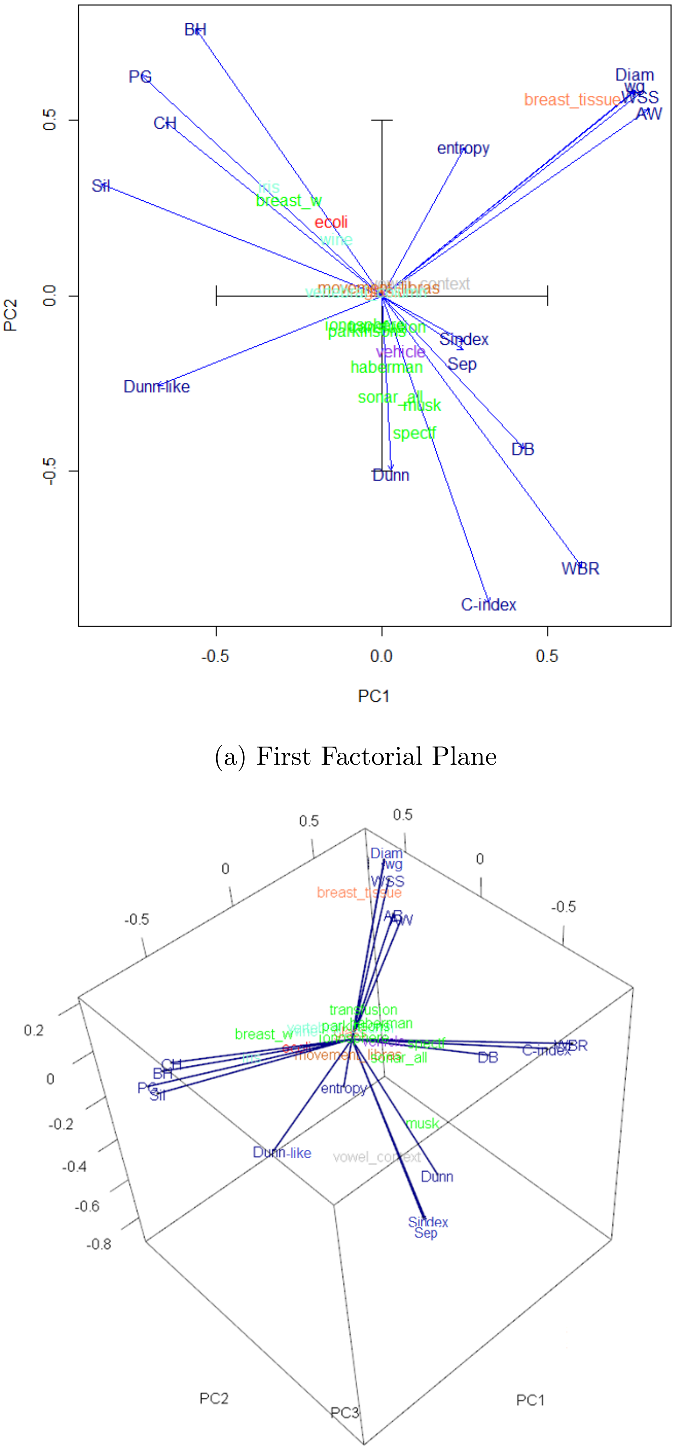
space. The analyzed datasets are represented by points. It can be seen that indexes place grouped over 5 directions both over the 1st factorial plane and the 1st factorial cube: On the one hand, all compactness indexes (Diameter(Δ), wg, $\overline{W}$, WSS) place together and orthogonally to most of the relational indexes (CH, Silhouettes, $\hat{\Gamma}$, BH, WBR, C-Index, DB); on a third direction, we can find most separation indexes (Separation (δ), Sindex); Entropy follows its own behavior; also the Dunn index behaves orthogonally of the other families, probably because it works with minimums and maximums and it is known to be less robust to noise [Bezdek and Pal, 1998]. Finally, Dunn-like seems to follow inverse association with compactness indexes.

Thus, using the multivariate information provided in this map one can determine whether the clusters of the datasets are more compact, more spherical or have big gaps. Also, if clusters seem to be overlapped or one can see which datasets have more separated data, which have at least two clusters nearer, or if they have one cluster sparser than others.

When dataset are plotted over this picture one can learn about the topology of the classes contained. For example, it seems that *breast_ tissue* has more compact classes that other datasets. While *musk* has more separated classes. Also, in *iris* data, are shown the best values in the indexes except in separation because it is well-known that 2 of the classes are non-linearly separable and, in fact, they have no a clear frontier between them.

In fact, as some packs of indexes behave similarly, according to the 5 directions they group on factorial space, one possibility could be to choose one representative index for every family and use the resulting reduced battery to evaluate datasets in a more efficient way.

However, in a real application the common situation to be faced is to get different partitions of the target dataset obtained by different clustering methods, rather than to have several datasets, to analyze and compare, which is a more common scenario for academic purposes. Thus, in a real application where several methods are used to cluster data, the global set of indexes might be used to learn about the characteristics of each partition and this can help to find the best one. In next section this possibility is illustrated with a real dataset.



(a) First Factorial Plane

(b) 3D projection over the first three factorial axes

Figure 7.3: Projection of the Firsts Factorial Axis

7.3.3 Application

The data used for this analysis corresponds with the data at the beginning of the intervention (see section 5.3). Cleaned data was clustered using the following 3 clustering methods:

- Ward's method. The resulting dendrogram has been cut in 6 and 12 classes according to Calinski - Harabasz Index.
- PAM, ξ has been selected using a meta-algorithm that checks several ξ's and propose the one optimizing the Calinski - Harabasz Index. Here the 3 best cuts are retained for comparison.
- Integrative Multiview Clustering. The partition P_o created in section 5.4.2.

Cluster validity indexes introduced before have been computed over the 6 partitions and shown in Table 7.11.

From Table 7.11, it is observed that each index provides a different ranking and there is not a consensus of which is the best partition. In fact, as said the indexes evaluate different aspects of a partition and can rank differently according to the classes topology.

It seems that the partition in 12 classes of the Hierarchical Clustering (HC12) is the best ranked in 8 of the indexes. The other partitions rank better or worse depending on the index.

HC12 and PAM8 seems to have more compact clusters in terms of the Within-cluster distances ($\overline{W}$), Sum of Squares of Within-cluster distances and Diameter (Δ). The two first ones indicate cohesion and the third one points to small clusters. P_o is the best regarding the widest gap index which evaluates whether the clusters are dense or rather contain gaps inside of the clusters. HC6 performs the worst in Diameter (Δ), this indicates that it contains at least one wider cluster than other partitions.

The separation indexes perform similarly for all partitions and show low values. This indicates that frontiers among classes are not so wide. HC12 and HC6 have small improvements in the Separation (δ) and Sindex indexes.

P_o and HC6 have better performed in those relational indexes which work with average separation (Dunn-like, Calinski-Harabasz). Notwithstanding, P_o has a bad position in relational indexes (Dunn, Silhouette, C-index) which work with minimum separations. This seems to indicate that separability of classes is not so high regarding inner variability. However, since it has better performance in indexes working with average separation (Dunn-like, Calinski-Harabasz) we can conclude that this partition includes some class which is really close to the boundary of some other cluster and this can distort indexes that work with minimums (Dunn, C-index).

HC12, HC6 and PAM8 are better ranked in those indexes which work with minimum separations (Dunn, Silhouette, C-index), except HC6 for Dunn. As Dunn has Diameter

Table 7.11: Cluster Validity Indexes computed over the 6 resulting partitions. The first three best values are highlighted in bold.

		Chaos		Compactness			Separation			Relational						
Dataset	Num Clusters	Entropy	Δ	wg	$\overline{W}$	WSS	$\overline{B}$	δ	Sindex	Dunn	Dunn-like	CH	$\hat{\Gamma}$	Silhouette	BH	WBR	C-Index
P_o	8	**1,939**	**0,355**	**0,234**	0,199	**1,645**	**0,232**	0,075	0,084	0,212	**0,902**	**5,281**	0,215	0,025	0,342	0,859	0,303
HC6	6	**1,737**	0,427	**0,237**	**0,191**	1,683	**0,234**	**0,095**	**0,102**	0,222	**0,896**	**7,036**	**0,292**	**0,064**	**0,441**	**0,817**	**0,251**
HC12	12	2,411	**0,336**	**0,237**	**0,175**	**1,276**	**0,232**	**0,095**	**0,098**	**0,281**	0,809	**6,142**	**0,287**	**0,080**	**0,579**	**0,754**	**0,197**
PAM6	6	**1,775**	0,366	**0,237**	0,201	1,818	**0,232**	0,084	0,091	**0,230**	0,887	5,274	0,204	0,033	0,315	0,867	0,313
PAM7	7	1,942	0,366	0,248	0,197	1,738	0,231	0,084	**0,093**	**0,230**	**0,946**	5,166	0,214	**0,040**	0,347	0,853	0,297
PAM8	8	2,048	**0,346**	**0,237**	**0,193**	**1,662**	0,231	0,084	**0,093**	**0,244**	0,856	5,111	**0,231**	0,037	**0,395**	**0,832**	**0,275**
Target		min	min	min	min	min	max	max	max	max	max	max	1	1	1	min	min

(Δ) in the denominator, this must probably be related with the worst ranked HC6 in Diameter (Δ).

In this application, the partitions HC6, HC12, PAM8 and P_o seem to be the better structured. HC6 is among the best ranked in 13 of 16 indexes, HC12 in 12, PAM8 in 10 and P_o in 8. From the joint analysis of all indexes it can be said that HC6 has compact and dense clusters with a really bigger cluster than other clusters; HC12 has compact, small and dense clusters; PAM8 has compact and small clusters and P_o has compact, small, dense and with smaller holes in the clusters and with some cluster really close to some other one. Partition P_o is often placed in between the hierarchical partitions in 12 clusters (HC12) and in 6 (HC6), and even improves in some indexes.

Thus, joint analysis of many indexes provides information about the topology of the clusters that helps to better evaluate these characteristics of the proposed Clustering. However, in most of the situations, like the present one, where some indexes perform better or worse for all the partitions it seems appropriated to consider additional criteria related to interpretability of the clusters to complement the structural information provided by the multivariate evaluation of the whole set of indexes. Thus, this joint evaluation using indexes would be used to discard PAM solutions and reduce the set of partitions shown to experts for interpretation.

7.3.4 Discussion

In this study, cluster validation through CVIs and indicators is faced. A joint multivariate evaluation of all indexes is proposed as a richer methodology than traditional and simple ranking according to indexes. The proposal provides information about the topology as well as the structure of the resulting partition.

Efforts have been inverted to express all considered indexes by means of a common notation (Section 2.5); this permitted a deep understanding of the indexes themselves, and to realize that most of them refer to some upper category that represent different characteristics of the clusters from a structural point of view: whether clusters are more compact or more sparse, or whether there are at least two classes that are too close, or whether it seems to be more or less overlapping among clusters. In fact, from this conceptual analysis we identify 4 categories of indexes: those meaning compactness, separation, relation between compactness and separation or chaos.

With the analysis of the relationships among those indexes by means of PCA, the indexes group depending on its behavior in front of the different datasets that have been analyzed and, they group accordingly to the conceptual classification provided in previous section. Except for Dunn index that has its own behavior and the average Between-cluster distances ($\overline{B}$) that seems to approach closer the compactness indexes.

For the reason of efficiency, this would allow choosing a reduced battery of 5 indexes to evaluate the structure of the partition, one referred to each relevant characteristic. One possibility would be: Entropy, Dunn, Diameter(Δ), Separation (δ) and Calinski-Harabasz indexes for evaluating a partitions while its topology is also understood.

Additionally, different partitions of the case study have been analyzed under the proposed approach. Table 7.11 provides an idea of the structure of these partitions and helps to select the more sound ones from a topological point of view to be presented to the experts for interpretation and further selection.

That is why it is recommended to decide whereas all indexes or a reduced representative battery is used depending on every particular application.

This research line has been motivated by the inherent difficulties of the cluster validation process. In most real applications, there is not prior information about the number of clusters or about the structural characteristics of the existent clusters.

The main contribution of this research is that joint multivariate vision of a complete set of indexes provides richer information about the topology and structure of the clusters than traditional ranking analysis.

A deep understanding of 17 CVIS and indicators found in the literature are really measuring is achieved through the analysis of their expressions under a common notation and a higher conceptual hierarchy of indexes emerged from this analysis based on the cluster characteristics targeted in each index. This explains why rankings obtained are different depending on the index for a set of datasets. PCA analysis confirms homogeneous behavior of indexes of same category in general trends.

Hence, from both the conceptual and the PCA analysis is possible to classify the indexes into 5 groups. Eventually a dataset can be assessed using all of them or a reduced battery containing a representative of each category. For example Entropy, Dunn, Diameter(Δ), Separation (δ) and Calinski-Harabasz.

The objective of this application is finding a clustering of the individuals being meaningful and useful for clinical practice. Several partitions were computed using different clustering algorithms. Using the cluster validity indexes, the topology of those partitions has been evaluated. And less sound partitions from the structural point of view could be identified and discarded.

7.4 Summary

This chapter contains two different lines to explore for the cluster interpretation and s study about the most common CVIs for cluster validation.

Section 7.1 present and alternative of finding the descriptors in a class when nested partitions exists. The main idea is to study the behavior of an attribute in the subclasses of the same superclass. In our application, since the final partition is nested in two partitions, this approach allows seeing the behavior of one superclass inside the other superclass.

On the another hand, Section 7.2 propose a complete different methodology for cluster interpretation using the frequent itemsets. From this work, the *Specificity Coefficient* is proposed to evaluate the itemsets inside of a class indicating if an itemset is predominating in this class. By now, this approach has only been tested with categorical attributes. However, this is an interesting future line of research.

A strategy to validate the resulting clusters based on CVIs is presented in Section 7.3. After an analysis of the behavior of the most common CVIs in front of different datasets, the CVIs are grouped accordingly to the conceptual classification: compactness, separation, relation between compactness and separation or chaos. Except for the Dunn index that has its own behavior.

With this chapter of the alternative lines of research, the contributions of this PhD Thesis are completed. In next chapter, the final conclusions are presented.

Chapter 8

Conclusions and Future Work

This research has been contextualized within the scope of nutritional pre-post intervention studies. This work has been motivated by the inherent difficulties of this type of studies when classical approaches were used, as referred in the Introduction (Chapter 1). This is an interesting research field since more resources are devoted to extract deeper knowledge for human genome and there is an increasing activity in the area.

Being an emergent field, there are not much databases available yet, and the current ones are limited by strict copy right properties or expensive costs. However, many clinical studies in the area are currently in progress in different countries and will be a relevant source of knowledge in near future. Providing a **new methodology to better extract the knowledge contained in these kinds of datasets** will contribute to a quicker development of new health approaches, like patient-centered medicine, where genetics of the subjects is taken into account to design personalized, more efficient and less risky treatments.

An analysis of the inner structure of this kind of problems has been done. As a result, the formalization presented is proposed in Section 3.1. The main idea is to **decompose the data matrix according to the different goals that attributes play in the study** in the sense of changing or not during the study or showing additive or multiplicative effects of the intervention.

In Chapter 2 we have revised the literature about pre-post studies, focusing on dietary intervention nutritional genomics studies. We have found out these features:

- The diet effect on certain metabolism (with a certain state) is local to the gene expression.
- Till now, models relating every single factor versus just a few subsets of genes previously preselected have been proposed. There is a lack of global models taking into account the complex multivariate interactions among both factors and genes.
- Commonly, in pre-post studies only numerical attributes are analyzed.

- For pre-post studies only traditional statistical techniques have been used. It should be noted that in only two studies, AI techniques were used. In both the AI method was a hierarchical clustering, but was just used as a postprocessing step to better organize results, and not for the analysis.

As a result of this review, we focus this PhD thesis on some lines we detected there is not much work done, even though we identified those lines as promising opportunities to improve the modeling and the learning in nutritional genomics and related areas. We refer to introduce AI techniques in the analysis. In particular, clustering and cluster interpretation methods are introduced as a first approach, to design a suitable methodology that properly integrates the different steps, and also, contributes in the field of automatic class interpretation.

In Section 3.2, the **Integrative Multiview Clustering (IMC) approach has been proposed**. On the one hand, IMC reduces the computational cost of analyzing high dimensional data. On the other hand, it provides more interpretable clusters. It decomposes the data into meaningful subsets that allow clustering subjects without mixing concepts to integrate the multiview results into a single partition that catches relationships between meaningful concepts. First, multiview principle is used to simplify the problem structure and to permit local complex models; then a further cross-clustering step integrates knowledge from all views in a single typology. Knowledge domain is used to identify the thematic blocks defining the views. The methodology is general and permits as many views as desired (which is not considered in other proposals as can be seen in Section 2.3.3).

The methodological proposal is being tested and applied to a case study in nutritional genomics. The clinical study has been understood and described in detail (see Section 5.1). The proposed methodology has been applied and results presented in Chapter 5. Being real, complex and quite big data (612 attributes), preprocessing had a heavy weight which is worthy of consideration. Section 5.2 gives details of the complete preprocessing process often dismissed in scientific documents, but which implies many a priori decisions and assumptions that we made explicit in this Section 5.2, to enable proper interpretation of final results. With preprocessing we refer specifically to the cleaning of empty, constant, redundant, null, erroneous and irrelevant attributes. Also, cleaning drug attributes and transforming them into a reduced set of less detailed indicators with sufficient information for the purposes of this work.

For the first part of the methodology, the datasets are created for the different views. Matrix χ that contains static biometric information about subjects like sex or height and matrix $\mathcal{Y}$ were split into 2 thematic bocks (see Section 5.3):

- $\mathcal{B}_1$: static and dynamic biometric markers like weight or pressure, health conditions and blood and urine analysis.

- $\mathcal{B}_2$: food habits and physical activity habits.

By crossing the multiview clustering and **integrating proper interpretation** tools in the process, a typology of persons has been found by integrating the intrinsic nature, their health baseline and diet and physical activity habits both at the beginning and at the end of the study.

The interpretation of the results of data mining analysis is an important task for guaranteeing a powerful support for further decision-making processes. In the specific context of clustering, it is important to understand the essence of the resulting classes, and to detect differential characteristics among classes. The field of cluster interpretation is rather neglected. In literature, most of the works are focused on the structural validation rather than in the meaning of the partition (see Section 2.5). Therefore, this post-process lacks automatic tools that enable a quick interpretation of the resulting classes (see Section 2.6).

In order to interpret these clusters, statistic hypothesis tests have been used so as to find the most relevant attributes in each class. The relevance of each attribute has been valued not only within the whole sample, but within each of the resulting clusters. The **specific *Test-Value* designed in the context of multivariate analysis** have been used for detecting the values of attributes with bigger contributions to the class formation. This can help to better understand the class and their particularities with respect to other classes. In this work, the *Test-Value* is presented as a powerful tool to interpret the classes, since this test subsumes the capacity of classical tests such Kruskal-Wallis, ANOVA or χ^2-Independence test. It is also seen than *Test-Value* provides richer information that avoids multiple comparison procedures traditionally required to identify the specific set of classes behaving differently from global trend (when global tests are significant). For this reason, the assessment through Kruskal-Wallis, ANOVA, or χ^2-Independence tests were eliminated for our interpretation procedures and directly assess the Test-Value. In fact, this *Test-Value* allows obtaining an initial characterization of the classes in the form of a profile (see Section 6.3).

However, the test was not always consistent and a **generalization of the *Test-Value* is proposed** in Section 3.3. A formal analysis of the reasons that produce this inconsistency showed that the effect of the different classes size might disturb the significance test performances, as significance is directly related with the variance of estimates, and inversely proportional to the sample size (see Section 6.4). On the other hand, **a Sensitive Analysis is proposed in order to evaluate the strength of the interpretations derived from the use of Test-Value**.

A **new interpretation methodology is introduced** in Section 3.6 based on the generalized *Test-Value* statistics to a better justified estimation of the variance, for the particular case when the two means compared correspond to non-independent samples, with a global mean and computed over a subsample of the same set of individuals.

The new methodology includes the new tests, as well as, the classification of descriptors according to their robustness degree to be used in the final interpretations with different levels of reliability according to robustness. Robustness is evaluated on the basis of sensitivity to smaller or larger changes in actual sample sizes. We refer to this evaluation as the **Sensitive Analysis**.

The particular case where two nested partitions have to be interpreted is tackled in this PhD. Thesis, from the point of view of guaranteeing the consistency of both interpretations. To this end, an accurate analysis of the different situations that might appear from the consistency point of view enabled the identification of four basic situations related with the inheritance of properties between parent and children classes: maintained irrelevance of a property in both partitions, maintained irrelevance in parent partition but relevant for some children, maintained relevance in parent partition and relevance for some children and inconsistency: the later being produced when a variable is significant in the parent class, but non-significant for all subclasses. A **second methodology of cluster interpretation is proposed** in Section 3.9 that provides consistent interpretations when nested partitions occur.

Also, the **introduction of the dynamic approach of the *Test-Value*** (see Section 3.3.2) allows characterizing the changes along the intervention in each of the found profiles.

The **adherence to the dietary intervention** has been evaluated in Section 5.7 concluding that, in this case, a dataset with higher granularity is needed in order to evaluate increase or decrease in the consumption of certain foods with more accuracy. However, our approaches (local to each class) gives preciser information than the one obtained under direct use of intervention groups (global). By the moment, the consumption of olive oil (either virgin or washed) is available. Using these new attribute, it has been shown that the methodology works properly and it is suitable to identify both the adherence to the intervention as possible irregularities in the adherence (see section 5.7.4).

A deeper analysis and **characterization of the Pre-Post Trajectories** has been performed in order to see how the different individuals have evolved during the study comparing both the initial and final state. The characterization of the resulting trajectories allows detecting the effect of the intervention.

Besides, one of the main interests in class interpretation is to discover local association among attributes specific for every class. That is why we propose to explore the possibility of using *frequent itemsets* to improve the interpretation process. Preliminary results are provided in this matter.

Frequent itemsets have been used for finding patterns suitable to define classes. We have defined a ***Specificity Coefficient* for determining the most characterizing attributes** within the clusters (see Equation 7.5 in Section 7.2).

Another research line was devoted to the **analysis the Cluster Validity Indexes** (CVI), commonly used for cluster validation purposes. The main result of our research was a multivariate analysis if 17 CVIs which provided a richer information about structure and topology than traditional ranking analysis.

Although initially the methodology was conceived to deal with nutritional intervention pre-post studies, it can be applied in other domains of health sciences such as in pharmacology or neurorehabilitation. Even, the flexibility of this methodology allows using it across other disciplines such as education, economics (policy effects, for example) and many other social sciences.

8.1 Contributions

The list of our research contributions can be summarized as follows:

1. Decompose the pre-post intervention study in such a way that local relationships between attributes and high-order interactions can be caught.
2. Modeling the intervention effect (pre-post) of diet locally for each of the class (clusters). Two effect models are considered: additive effect (for which difference are used) and multiplicative effect (for which ratios are more indicated).
3. Introduction of the Integrative Multiview Clustering.
4. Importation of the *Test-Value* for Cluster Interpretation
5. Generalization of the *Test-Value* for being less sensitive to the class sizes.
6. Dynamic approach of the *Test-Value* to determine the effects of the intervention.
7. Introduction of the Sensitive Analysis for analyzing the robustness of the descriptors.
8. Proposal of the Cluster Interpretation methodology CI-IMS, thus yielding an automatic description of the classes.
9. Treatment of possible inconsistencies between the interpretations of nested partitions.
10. Introduction of the Cluster Interpretation methodology NCI-IMS for nested partitions.
11. Characterization of the Pre-Post Trajectories. This characterization allows evaluating the adherence to the intervention and analyzing the intervention effects.
12. Local approach of the *Test-Value.*
13. Use the Frequent itemsets with the definition of an optimal criterium (*Specificity Coefficient*).
14. Accurate analysis of the CVIs proposing a combination of indexes to evaluate a clustering.

15. The last and the most important contribution – being the result of having done all the ones before – is the proposal of a new methodology for dealing with complex pre-post intervention studies as it is detailed in Chapter 3.

8.2 Future Work

By now, the main future research we are envisioning is described in this section.

The first step will be to develop the Intelligent Decision Support System prototype as a guidance tool as it is described in Section 4.3. This tool will include the proposed methodology for the pre-post intervention studies. This interactive tool will automatically generate the Trajectory Map, which allows interpreting the behavior of the individuals according to the assigned intervention. In addition, by clicking on the name of a class, its profile will be shown, and clicking on a trajectory, the adherence to the intervention or the intervention effects will be depicted.

As it is stated in Section 5.7.4, the first improvement of this thesis would be to analyze the adherence to the intervention with a more refined data where quantities of food are not discretized, in order to see how the dietary habits have changed along the study.

In addition, the building of an artificial dataset with known properties could be used for demonstrating the power of the proposed methodology. Also, it is very interesting to test the proposed methodology with other pre-post intervention studies when data becomes available.

Other research lines to explore are related with a possible improvement of the cluster interpretation methodology. For instance, introducing the Local Test-Value in the Cluster Interpretation Methodology with nested partitions as this approach let to see the behavior of one partition inside other partition. Other possibility is exploring the itemsets for cluster interpretation and also, that intemsets deal with numerical attributes. Besides, other line to explore would be searching for the less frequent itemsets in a class. An itemset with a very low *Specificity Coefficient* could be interesting for describing the class by its absence.

Regarding the Integrative Multiview clustering, it can be interesting the automatic generation of the thematic blocks using semantic analysis or ontologies. Due to the fact that the Integrative Multiview Clustering can work with other clustering techniques (different to the hierarchical clustering), the system could include a methodology for a prior selection of the underlying clustering technique. This methodology could be based on the structural validation using one CVI of the 5 types founded in Section 7.3.

Finally, the comparison of the profiles can be refined taking profit that our profiles contains the robustness strength of each descriptor, and therefore, different degrees of differences can be introduced.

Chapter 9

Publications

Related Publications:

Journal Papers:

- B. Sevilla-Villanueva, K. Gibert, M. Sànchez-Marrè.*A Methodological Framework to Understand Nutritional Patterns and Adherence to Diet in Intervention Studies based on Data Mining Methods.* Submitted to Artificial Intelligence in Medicine.s

- K. Gibert, B. Sevilla-Villanueva, M. Sànchez-Marrè. *The role of significance tests in consistent interpretation of nested partitions.* Journal of Computational and Applied Mathematics, 2015.

  doi:10.1016/j.cam.2015.01.031 ISSN: 0377-0427

  Science Citation Index: 1.104 Impact Factor: 1.266 (QI)

Chapters in collections:

- B. Sevilla-Villanueva, K. Gibert, M. Sànchez-Marrè. *Identifying nutritional patterns through Integrative Multiview Clustering.* Frontiers in artificial intelligence and applications, 277:185-194, IOS Press 2015.

  isbn:97-1-61499-577-7

- B. Sevilla Villanueva, K. Gibert, M. Sánchez-Marrè, M. *Post-processing the Class Panel Graphs: towards understandable patterns from data.* Frontiers in artificial intelligence and applications, 256:215-224, IOS Press 2013.

  isbn: 978-1-61499-319-3

  DOI:10.3233/978-1-61499-320-9-215

Conference Papers:

- B. Sevilla-Villanueva, K. Gibert, M. Sànchez-Marrè. *The Role of Statistical Tests on Cluster Interpretation.* Mathematical Models in Engineering & Human Behaviour, 134-137, 2014.
- B. Sevilla-Villanueva, K. Gibert, M. Sanchez-Marre, M. *Clustering and interpretation on real nutritional data.* In Conference VII Simposio de Teoría y Aplicaciones de Minería de Datos, 2013.

Other Publications:

Chapters in collections:

- B. Sevilla-Villanueva, M. Sànchez-Marrè, T.V. Fischer. *Estimation of Machine Settings for Spinning of Yarns ? New Algorithms for Comparing Complex Structures.* In Case-Based Reasoning Research and Development. Springer International Publishing, 435-449, 2014.

  DOI: 10.1007/978-3-319-11209-1_31 Print ISBN:978-3-319-11208-4
- B. Sevilla, M. Sànchez-Marrè. *Case-Based Reasoning Applied to Textile Industry Processes.* Procc. of 20th International Conference on Case-Based Reasoning (ICCBR'2012) "Lecture notes in computer science", 2012, 7466:428-442.

  doi: 10.1007/978-3-642-32986-9_32

Conference Papers:

- M. Sànchez-Marrè, K. Gibert, R. K. Vinayagam, B. Sevilla-Villanueva, *Evolutionary Computation and Case-Based Reasoning Interoperation in IEDSS through GESCONDA.* Proceedings of the 7th International Congress on Environmental Modelling and Software, 2014.
- K. Gibert, M. Sánchez-Marrè, B. Sevilla. *Tools for environmental data mining and intelligent decision support.* 6th International Congress on Environmental Modelling & Software (iEMSs 2012) iEMSs'2012 Procceedings, 1726-1734. R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (eds.) Leipzig, Germany, Julio 2012.

  ISBN: 978-88-9035-742-8
- M. Sànchez-Marrè, K. Gibert and B. Sevilla (2010). *Evolving GESCONDA to an Intelligent Decision Support Tool.* 5th International Congress on Environmental Modelling and Software (iEMSs'2010) iEMSs'2010 Procceedings, 3:2015-2024. Ottawa, Canada. 5-8 de Julio de 2010

  ISBN 978-88-903574-1-1

- B. Sevilla and M. Sànchez-Marrè (2010). *Providing Intelligent Decision Support Systems with Flexible Data-Intensive Case-Based Reasoning.* 5th International Congress on Environmental Modelling and Software (iEMSs'2010) iEMSs'2010 Procceedings, 3: 2063-2072. Ottawa, Canada. ISBN 978-88-903574-1-1

References

[Abdullin and Nasraoui, 2012] Abdullin, A. and Nasraoui, O. (2012). Clustering heterogeneous data sets. In *Web Congress (LA-WEB), 2012 Eighth Latin American*, pages 1–8. IEEE. 29

[Afman and Müller, 2006] Afman, L. and Müller, M. (2006). Nutrigenomics: From molecular nutrition to prevention of disease. *Journal of the American Dietetic Association*, 106(4):569–576. 7, 15

[Agrawal et al., 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216. 222

[Alizadeh et al., 2000] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511. 30

[Anand Rajaraman and Ullman, 2011] Anand Rajaraman, J. L. and Ullman, J. D. (October 2011). Frequent itemsets. In *Mining of Massive Datasets*, pages 183–220. Cambridge University Press. 220

[Anderberg, 1973] Anderberg, M. R. (1973). *Cluster Analysis for aplications*. Academic Press. 34

[Arbelaitz et al., 2013] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., and Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256. 9, 36, 37, 38, 230

[Ayadi et al., 2012] Ayadi, W., Elloumi, M., and Hao, J.-K. (2012). Bicfinder: a biclustering algorithm for microarray data analysis. *Knowl. Inf. Syst.*, 30(2):341–358. 30, 33

[Baker and Hubert, 1975] Baker, F. B. and Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349):31–38. 42

[Balanza, 2007] Balanza, B. P. (2007). *Técnicas de Evaluación en Algoritmos de Biclustering sobre Datos de Expresión Genómica*. PhD thesis, Universidad de Sevilla. 16

[Barnard et al., 2001] Barnard, K., Duygulu, P., and Forsyth, D. (2001). Clustering art. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–434–II–441 vol.2. 44

[Bartlett and Stirling, 2003] Bartlett, J. and Stirling, D. (2003). A short history of the polymerase chain reaction. In Bartlett, J. and Stirling, D., editors, *PCR Protocols*, volume 226 of *Methods in Molecular Biology™*, pages 3–6. Humana Press. 16

[Benfey and Mitchell-Olds, 2008] Benfey, P. N. and Mitchell-Olds, T. (2008). From genotype to phenotype: systems biology meets natural variation. *Science*, 320(5875):495–497. 13

[Benzécri, 1973] Benzécri, J. (1973). *L'analyse des données*. Paris: Dunod. Tome 1: La Taxinomie, Tome 2: L'analyse des correspondances. First Edition. 36, 46

[Berge et al., 2000] Berge, K. E., Tian, H., Graf, G. A., Yu, L., Grishin, N. V., Schultz, J., Kwiterovich, P., Shan, B., Barnes, R., and Hobbs, H. H. (2000). Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent abc transporters. *Science*, 290(5497):1771–1775. 15

[Berkhin, 2006] Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer. 27

[Bezdek and Pal, 1998] Bezdek, J. C. and Pal, N. R. (1998). Some new indexes of cluster validity. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 28(3):301–315. 36, 37, 38, 41, 235

[Bickel and Scheffer, 2004] Bickel, S. and Scheffer, T. (2004). Multi-view clustering. In *Proceedings of the IEEE international conference on data mining*, volume 4, pages 19–26. 29, 197

[Blatt et al., 1996] Blatt, M., Wiseman, S., and Domany, E. (1996). Superparamagnetic Clustering of Data. *Physical Review Letters*, 76:3251–3254. 33

[Borgelt, 2003] Borgelt, C. (2003). Efficient implementations of apriori and eclat. In *Workshop Frequent Item Set Mining Implementations (FIMI 2003, Melbourne, FL, USA)*, Aachen, Germany. CEUR Workshop Proceedings 90. 222

[Brun et al., 2007] Brun, . M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., and Dougherty, E. (2007). Model-based evaluation of clustering validation measures. *Pattern Recognition*, 40(3):807–824. 37, 38, 230

[Caliński and Harabasz, 1974] Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27. 41

[Camargo et al., 2011] Camargo, A., Ruano, J., Fernández, J. M., Parnell, L. D., Jiménez, A., Santos-González, M., Marín, C., Pérez-Martínez, P., Uceda, M., López-Miranda, J., and Pérez-Jiménez, F. (2011). Virgin olive oil phenolic compounds interact with cellular signalling pathways. *Revista:clinica e investigacion arteriosclerosis Clínica e investigación en Arteriosclerosis*, 23(6):262–8. 19, 23

[Castañer et al., 2012] Castañer, O., Covas, M.-I., Khymenets, O., Nyyssonen, K., Konstantinidou, V., Zunft, H.-F., de la Torre, R., Muñoz-Aguayo, D., Vila, J., and Fitó, M. (2012). Protection of ldl from oxidation by olive oil polyphenols is associated with a downregulation of cd40-ligand expression and its downstream products in vivo in humans. *The American Journal of Clinical Nutrition*, 95(5):1238–1244. 19, 23

[Cecere and Abreu, 2010] Cecere, W. and Abreu, D. A. (2010). A method for improving list building: Cluster profiling. In *Proceedings of the Survey Research Methods Section, American Statistical Association*. 44

[CENG, 2015] CENG (2015). Center of excellence for nutritional genomics (ceng) at the university of california, davis -. `http://nutrigenomics.ucdavis.edu/`. Accessed: 2015-07-31. 18

[Chae et al., 2001] Chae, Y. M., Ho, S. H., Cho, K. W., Lee, D. H., and Ji, S. H. (2001). Data mining approach to policy analysis in a health insurance domain. *International Journal of Medical Informatics*, 62(2–3):103 – 111. 19, 21

[Cheeseman and Oldford, 2012] Cheeseman, P. and Oldford, R. W. (2012). *Selecting models from data: artificial intelligence and statistics IV*, volume 89. Springer Science & Business Media. 36

[Cheng and Church, 2000] Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103. AAAI Press. 30, 32

[Clarke et al., 2009] Clarke, B., Fokoue, E., and Zhang, H. H. (2009). *Principles and theory for data mining and machine learning*. Springer Science & Business Media. 28

[Clement et al., 2004] Clement, K., Viguerie, N., Poitou, C., Carette, C., Pelloux, V., Curat, C., Sicard, A., Rome, S., Benis, A., Zucker, J., et al. (2004). Weight loss regulates inflammation-related genes in white adipose tissue of obese subjects. *The FASEB Journal*, 18(14):1657–1669. 19, 25

[Coltell i Simon, 2004] Coltell i Simon, s. (2004). *Integración de la Bioinformática en la Investigación Genómica Cardiovascular: Aplicaciones en el Framingham Heart Study*. PhD thesis, En la Universitat Jaume I (España). 19, 22

[Corella et al., 2009] Corella, D., Peloso, G., Arnett, D. K., Demissie, S., Cupples, L. A., Tucker, K., Lai, C. Q., Parnell, L. D., Coltell, O., Lee, Y. C., and Ordovas, J. M. (2009). Apoa2, dietary fat, and body mass index: replication of a gene-diet interaction in 3 independent populations. *Arch Intern Med*, 169(20):1897–1906. 18

[Cortez and Embrechts, 2012] Cortez, P. and Embrechts, M. J. (2012). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 225(0):1—-17. 44

[Covas et al., 2006] Covas, M.-I., Nyyssönen, K., Poulsen, H. E., Kaikkonen, J., Zunft, H.-J. F., Kiesewetter, H., Gaddi, A., de la Torre, R., Mursu, J., Bäumler, H., Nascetti, S., Salonen, J. T., Fitó, M., Virtanen, J., and Marrugat, J. (2006). The effect of polyphenols in olive oil on heart disease risk factors: A randomized trial. *Annals of Internal Medicine*, 145(5):333–341. 23

[Dahl, 2013] Dahl, D. B. (2013). *xtable: Export tables to LaTeX or HTML*. R package version 1.7-1. 96

[Dahlman et al., 2005] Dahlman, I., Linder, K., Arvidsson Nordström, E., Andersson, I., Lidén, J., Verdich, C., Sørensen, T. I., Arner, P., and Nugenob (2005). Changes in adipose tissue gene expression with energy-restricted diets in obese women. *The American Journal of Clinical Nutrition*, 81(6):1275–1285. 19, 25

[Davies and Bouldin, 1979] Davies, D. L. and Bouldin, D. W. (1979). Cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):95–104. 43

[de Lorenzo, 2012] de Lorenzo, D. (2012). Perspectivas presentes y futuras de la nutrigenómica y la nutrigenética en la medicina preventiva. *Nutrición Clínica y Dietética Hospitalaria*, 32. 17

[DeRisi et al., 1996] DeRisi, J., Penland, L., Brown, P., Bittner, M., Meltzer, P., Ray, M., Chen, Y., Su, Y., , and Trent, J. (1996). Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14:457–460. 16

[Dimitriadou et al., 2002] Dimitriadou, . E., Dolnicar, S., and Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary datasets. *Psychometrika*, 67:137–159. 36, 37, 38, 40, 230

[D'Orazio, 2012] D'Orazio, M. (2012). *StatMatch: Statistical Matching*. R package version 1.0.5. 35, 95

[Dubes, 1987] Dubes, . R. (1987). How many clusters are best? - an experiment. *Pattern Recognition*, 20(6):645?–663. 36, 37, 38, 230

[Duda et al., 2012] Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons. 26, 230

[Dunn, 1974] Dunn, J. C. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104. 40, 41

[Eisen et al., 1998] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868. 30

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*. 194

[Estruch et al., 2013] Estruch, R., Ros, E., Salas-Salvadó, J., Covas, M.-I., Corella, D., Arós, F., Gómez-Gracia, E., Ruiz-Gutiérrez, V., Fiol, M., Lapetra, J., Lamuela-Raventos, R. M., Serra-Majem, L., Pintó, X., Basora, J., Muñoz, M. A., Sorlí, J. V., Martínez, J. A., and Martínez-González, M. A. (2013). Primary prevention of cardiovascular disease with a mediterranean diet. *New England Journal of Medicine*, 368(14):1279–1290. PMID: 23432189. 99

[Farina et al., 2011] Farina, E., Kiel, D., Roubenoff, R., Schaefer, E., Cupples, L., and Tucker, K. (2011). Dietary intakes of arachidonic acid and alpha-linolenic acid are associated with reduced risk of hip fracture in older adults. *The Journal of nutrition*, 141(6):1146–1153. 19, 22

[Fayyad et al., 1996] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. 38

[FDA, 2015] FDA (2015). U.s. food and drug administration. `http://www.fda.gov/`. Accessed: 2015-07-31. 17

[Fenech, 2008] Fenech, M. (2008). The human genome, nutrigenomics and nutrigenetics. *Innovation: Management, Policy & Practice*. 14

[Fraley and Raftery, 2002] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631. 95, 194

[Framingham Heart Study, 2015] Framingham Heart Study (2015). A project of the national heart, lung and blood institute (nhlbi). `http://www.framinghamheartstudy.org/`. Accessed: 2015-07-31. 6, 18

[Franklin and Gosling, 1953] Franklin, R. E. and Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature*, 171:740–741. 16

[Gasch and Eisen, 2002] Gasch, A. P. and Eisen, M. B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11):research0059.1–research0059.22. 30, 31

[Getz et al., 2000] Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences*, 97(22):12079–12084. 30, 33

[Geurts et al., 2003] Geurts, K., Brijs, T., and Wets, G. (2003). Clustering and profiling traffic roads by means of accident data. In *Proceeding of the European Transport Conference (ETC), Strasbourg, France*. 44

[Ghouila et al., 2009] Ghouila, A., Yahia, S. B., Malouche, D., Jmel, H., Laouini, D., Guerfali, F. Z., and Abdelhak, S. (2009). Application of multi-som clustering approach to macrophage gene expression analysis. *Infect Genet Evol*, 9(3):328–336. 30, 31

[Gibert, 2014] Gibert, . K. (2014). Automatic generation of classes interpretation as a bridge between clustering and decision making. *International Journal of Multicriteria Decision Making.*, 4(2):154–182. 9, 44

[Gibert, 1994] Gibert, K. (1994). *L'us de la Informació Simbòlica en l'Automatització del Tractament Estadístic de Dominis Poc Estructurats*. phd. thesis., UPC, Barcelona, Spain. 34

[Gibert and Annicchiarico, 2003] Gibert, K. and Annicchiarico, R. e. a. (2003). Kdd on functional disabilities using clustering based on rules on who-das ii. In *ITI 03.*, pages 181–186, Croatia. 36

[Gibert and Conti, 2014] Gibert, K. and Conti, D. (2014). On the understanding of profiles by means of post-processing techniques: an application to financial assets. *International Journal of Computer Mathematics*, (ahead-of-print):1–14. 9

[Gibert et al., 2012] Gibert, K., Conti, D., and Vrecko, D. (2012). Assisting the end-user in the interpretation of profiles for decision support. an application to wastewater treatment plants. *Environmental Engineering and Management Journal*, 11(5):931–944. 44

[Gibert and Cortés, 1992] Gibert, K. and Cortés, U. (1992). KLASS: Una herramienta estadística para la creación de prototipos en dominios poco estructurados. In *IBERAMIA-92.*, pages 483–497, México. Noriega Eds. 34, 36

[Gibert and Cortés, 1997] Gibert, K. and Cortés, U. (1997.). Weighing quantitative and qualitative variables in clustering methods. *Mathware and Soft Computing*, 4(3):251–266. 36

[Gibert and Cortés, 1998] Gibert, K. and Cortés, U. (1998.). Clustering based on rules and knowledge discovery in ill-structured domains. *Computación y Sistemas.*, 1(4):213–227. 36

[Gibert and García, 2008] Gibert, K. and García, A. (2008). Posibilidades de aplicación de minería de datos para el descubrimiento de conocimiento a partir de la práctica clínica. *Tecnologías Aplicadas al Proceso Neurorrehabilitador : Estrategias para Valorar su Eficacia.* 18, 21

[Gibert et al., 2010a] Gibert, K., García-Alonso, C., and Salvador-Carulla, L. (2010a). Integrating clinicians, knowledge and data: expert-based cooperative analysis in healthcare decision support. *Health research policy and systems BioMed Central*, 8(28):28. 46

[Gibert et al., 2008a] Gibert, K., García-Rudolph, A., and Rodríguez-Silva, G. (2008a). The role of kdd support-interpretation tools in the conceptualization of medical profiles: An application to neurorehabilitation. *Acta Inform Med.* 9, 44, 45

[Gibert et al., 2008b] Gibert, K., García-Rudolph, A., García-Molina, A., Roig-Rovira, T., Bernabeu, M., and Tormos, J. M. (2008b). Knowledge discovery on the response to neurorehabilitation treatment of patients with traumatic brain injury through an ai&stats and graphical hybrid methodology. In *Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence*, pages 170–177, Amsterdam, The Netherlands, The Netherlands. IOS Press. 21

[Gibert et al., 2005] Gibert, K., Nonell, R., Velarde, J., and Colillas, M. M. (2005). Knowledge discovery with clustering: impact of metrics and reporting phase by using klass. *Neural Network World*, 4:319–326. 45

[Gibert and Roda, 2000] Gibert, K. and Roda, I. (2000). Identifying characteristic situations in wastewater treatment plants. In *Workshop BESAI (ECAI)*, volume 1, pages 1–9. 36

[Gibert et al., 2010b] Gibert, K., Rodríguez-Silva, G., and Rodríguez-Roda, I. (2010b). Knowledge discovery with clustering based on rules by states: a water treatment application. *Environ. Modell. Softw.*, 25(6):712–?723. 46

[Gibert et al., 2013] Gibert, K., Rodríguez-Silva, G., and Annicchiarico, R. (2013). Post-processing: Bridging the gap between modelling and effective decision-support. the profile assessment grid in human behaviour. *Mathematical and Computer Modelling*, 57(7-8):1633–1639. 9, 44

[Gibert et al., 2015] Gibert, K., Sevilla-Villanueva, B., and Sànchez-Marrè, M. (2015). The role of significance tests in consistent interpretation of nested partitions. *Journal of Computational and Applied Mathematics*. 59, 206, 213

[Gibert and Sonicki, 1999a] Gibert, K. and Sonicki, Z. (1999a). Classification based on rules and medical research. *Journal of Applied Stochastic Models and Data Analysis, formerly JAMSDA*, 15(3):319–24. 10, 46

[Gibert and Sonicki, 1999b] Gibert, K. and Sonicki, Z. (1999b). Classification based on rules and thyroids dysfunctions. *Applied Stochastic Models in Business and Industry*, 15(4):319–324. 36

[Gibert and Tormos, 2014] Gibert, K. and Tormos, J. M. (2014). From big data to decisional knowledge. *Pan European Networks: Science and Technology*, 12:125. 9

[Gibert et al., 2014] Gibert, K., Valls, A., and Batet, M. (2014). Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering. *Knowledge and Information Systems*, 40(3):559–593. 49

[Golub et al., 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537. 17

[Goodman and Kruskal, 1954] Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764. 42

[Gordon, 1999] Gordon, A. D. (1999). *Classification.* 2nd ed. Chapman and Hall/CRC. 38, 43

[Gowda and Diday, 1992] Gowda, K. C. and Diday, E. (1992). Symbolic clustering using a new similarity measure. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(2):368–378. 35

[Gower, 1971] Gower, J. (1971). A General coefficient if similarity and some of its properties. *Biometrics*, 27:857–874. 35

[Gremalschi and Altun, 2008] Gremalschi, S. and Altun, G. (2008). Mean squared residue based biclustering algorithms. In *Proceedings of the 4th international conference on Bioinformatics research and applications*, ISBRA'08, pages 232–243, Berlin, Heidelberg. Springer-Verlag. 30, 32

[Hahsler et al., 2012] Hahsler, M., Buchta, C., Gruen, B., and Hornik, K. (2012). *arules: Mining Association Rules and Frequent Itemsets*. R package version 1.0-12. 96

[Hahsler et al., 2005] Hahsler, M., Gruen, B., and Hornik, K. (2005). arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15):1–25. 96

[Halkidi et al., 2001] Halkidi, . M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *J. Intelligent Information Systems*, 17(2):107–145. 36, 38, 41, 230

[Halkidi et al., 2002a] Halkidi, . M., Batistakis, Y., and Vazirgiannis, M. (2002a). Cluster validity methods: part i. *ACM Sigmod Record*, 31(2):40–45. 36, 38, 230

[Halkidi et al., 2002b] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002b). Clustering validity checking methods: part ii. *ACM Sigmod Record*, 31(3):19–27. 36, 38, 39, 230

[Halkidi and Vazirgiannis, 2002] Halkidi, M. and Vazirgiannis, M. (2002). Clustering validity assessment using multi representatives. *In Proceedings of SETN Conference*. 40

[Han et al., 2011] Han, J., Kamber, M., and Pei, J. (2011). *Data mining: concepts and techniques: concepts and techniques*. Elsevier. 49

[Hartigan, 1972] Hartigan, J. A. (1972). Direct Clustering of a Data Matrix. *Journal of The American Statistical Association*, 67:123–129. 32

[Haughton et al., 2009] Haughton, D., Legrand, P., and Woolford, S. (2009). Review of three latent class cluster analysis packages: Latent gold, polca, and mclust. *American Statistician*, 63(1):81–91. 44

[He and Hui, 2009] He, Y. and Hui, S. C. (2009). Exploring ant-based algorithms for gene expression data analysis. *Artificial Intelligence in Medicine*, 47(2):105 – 119. 30, 31

[Hennig, 2013a] Hennig, C. (2013a). *fpc: Flexible procedures for clustering*. R package version 2.1-5. 41, 95, 230

[Hennig, 2013b] Hennig, C. (2013b). How many bee species? a case study in determining the number of clusters. *in Proceedings of GfKl-2012*, Hildesheim. 38, 39, 40

[Hennig and Liao, 2010] Hennig, C. and Liao, T. F. (2010). Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification. Technical report, Technical report. 39, 40

[Houle et al., 2010] Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nat Rev Genet*, 11(12):855–866. 13

[Hubert and Levin, 1976] Hubert, L. J. and Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83(6):1072–1080. 43

[Hulshof et al., 1992] Hulshof, K., Wedel, M., Löwik, M., Kok, F., Kistemaker, C., Hermus, R., Ten Hoor, F., and Ockhuizen, T. (1992). Clustering of dietary variables and other lifestyle factors (dutch nutritional surveillance system). *Journal of Epidemiology and Community Health*, 46(4):417–424. 29

[Human Genome Project, 2015] Human Genome Project (2015). Human genome project (hgp). http://web.ornl.gov/sci/techresources/Human_Genome/index.shtml. Accessed: 2015-07-31. 17

[Ichino and Yaguchi, 1994] Ichino, M. and Yaguchi, H. (1994). Generalized Minkowski Metrics for Mixed feature-type data analysis. *IEEE Tr SMC*, 22(2):146–153. 36

[International Human Genome Sequencing Consortium et al., 2004] International Human Genome Sequencing Consortium et al. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945. 17

[ISNN, 2015] ISNN (2015). International society of nutrigenetics / nutrigenomics. http://www.nutritionandgenetics.org/. Accessed: 2015-07-31. 18

[Jaimungal et al., 2011] Jaimungal, S., Wehmeier, K., Mooradian, A. D., and Haas, M. J. (2011). The emerging evidence for vitamin d-mediated regulation of apolipoprotein a-i synthesis. *Nutr Res*, 31(11):805–812. 18

[Jain, 2010] Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666. 28

[J.D. and F.H.C, 1953] J.D., W. and F.H.C, C. (1953). A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738. 16

[JINGO, 2015] JINGO (2015). The joint irish nutrigenomics organisation (jingo). http://www.ucd.ie/jingo/. Accessed: 2015-07-31. 18

[Kaufman and Rousseeuw, 1987] Kaufman, L. and Rousseeuw, P. (1987). Clustering by means of medoids. *Statistical Data Analysis Based on the L1–Norm and Related Methods*, pages 405–416. 28

[Khymenets et al., 2009] Khymenets, O., Fitó, M., Covas, M.-I., Farré, M., Pujadas, M.-A., Muñoz, D., Konstantinidou, V., and de la Torre, R. (2009). Mononuclear cell transcriptome response after sustained virgin olive oil consumption in humans: An exploratory nutrigenomics study. *Omics: A Journal of Integrative Biology*, 13(1). 19, 23

[Kim and Ramakrishna, 2005] Kim, M. and Ramakrishna, R. (2005). New indices for cluster validity assessment. *Pattern Recognition Letters*, 26(15):2353–2363. 38

[Kimokoti et al., 2012] Kimokoti, R. W., Newby, P. K., Gona, P., Zhu, L., Campbell, W. R., D'Agostino, R. B., and Millen, B. E. (2012). Stability of the framingham nutritional risk score and its component nutrients over 8 years: the framingham nutrition studies. *Eur J Clin Nutr*, 66(3):336–344. 18

[Kofler et al., 2012] Kofler, B. M., Miles, E. A., Curtis, P., Armah, C. K., Tricon, S., Grew, J., Napper, F. L., Farrell, L., Lietz, G., Packard, C. J., Caslake, M. J., Mathers, J. C., Williams, C. M., Calder, P. C., and Minihane, A. M. (2012). Apolipoprotein e genotype and the cardiovascular disease risk phenotype: impact of sex and adiposity (the fingen study). *Atherosclerosis*, 221(2):467–470. 18

[Konstantinidou et al., 2010a] Konstantinidou, V., Covas, M.-I., Muqoz-Aguayo, D., Khymenets, O., de la Torre, R., Saez, G., Tormos, M. d. C., Toledo, E., Marti, A., Ruiz-Gutiirrez, V., Ruiz Mendez, M. V., and Fito, M. (2010a). In vivo nutrigenomic effects of virgin olive oil polyphenols within the frame of the mediterranean diet: a randomized controlled trial. *FASEB J*, 24(7):2546–57. 19, 23, 99, 182

[Konstantinidou et al., 2009] Konstantinidou, V., Khymenets, O., Covas, M.-I., de la Torre, R., Anglada, R., Dopazo, A., and Covas, M.-I. (2009). Characterization of human gene expression changes after olive oil ingestion: an exploratory approach. *Folia Biologica*, 55:77–83. 23

[Konstantinidou et al., 2010b] Konstantinidou, V., Khymenets, O., Fito, M., de la Torre, R., Muñoz-Aguayo, D., Anglada, R., Faree, M., and Fito, M. (2010b). Time course of changes in the expression of insulin sensitivuy-related genes after an acute load of virgin olive oil. *Omics: A Journal of Integrative Biology*, 13(4). 19, 23

[Kriegel et al., 2009] Kriegel, H.-P., Kröger, P., and Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data*, 3(1):1:1–1:58. 30, 33

[Kruskal and Wallis, 1952] Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *the American Statistical Association*, 47(260):583–621. 47

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921. 17

[Lausted et al., 2004] Lausted, C., Dahl, T., Warren, C., King, K., Smith, K., Johnson, M., Saleem, R., Aitchison, J., Hood, L., and Lasky, S. (2004). Posam: a fast, flexible, opensource, inkjet oligonucleotide synthesizer and microarrayer. *Genome Biology*, 5(8). 16

[Lebart et al., 1990] Lebart, L., Morineau, A., and Fénelon, J.-P. (1990). *Traitement statistique des données*. Dunod, Paris. 34

[Lebart et al., 2000] Lebart, L., Piron, M., and Morineau, A. (2000). *Statistique exploratoire multidimensionnelle*. Dunob, 3 edition. 10, 45, 47, 48, 58

[Lee et al., 2011] Lee, Y.-C., Lai, C.-Q., Ordovas, J. M., and Parnell, L. D. (2011). A database of gene-environment interactions pertaining to blood lipid traits, cardiovascular disease and type 2 diabetes. *Journal of data mining in genomics proteomics*, 2(1):1–8. 18

[Leisch, 2002] Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In Härdle, W. and Rönz, B., editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9. 94

[Li et al., 1993a] Li, B., Taylor, P. R., Li, J. Y., Dawsey, S. M., Wang, W., Tangrea, J. A., Liu, B. Q., Ershow, A. G., Zheng, S. F., and Fraumeni, J. F. (1993a). Linxian nutrition intervention trials. design, methods, participant characteristics, and compliance. *Ann Epidemiol*, 3(6):577–585. 19, 21

[Li and Shafto, 2011] Li, D. and Shafto, P. (2011). Bayesian hierarchical cross-clustering. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), Geoffrey Gordon, David Dunson, and Miroslav Dudk, eds. JMLR W&CP*, volume 15, pages 443–451. 29

[Li et al., 1995] Li, J., Li, B., Blot, W. J., and Taylor, P. R. (1995). Preliminary report on the results of nutrition prevention trials of cancer and other common diseases among residents in linxian, china. *Chin Med J (Engl)*, 108(10):780–780. 19, 21

[Li et al., 1993b] Li, J. Y., Taylor, P. R., Li, B., Dawsey, S., Wang, G. Q., Ershow, A. G., Guo, W., Liu, S. F., Yang, C. S., and Shen, Q. (1993b). Nutrition intervention trials in linxian, china: multiple vitamin/mineral supplementation, cancer incidence, and disease-specific mortality among adults with esophageal dysplasia. *J Natl Cancer Inst*, 85(18):1492–1498. 19, 21

[Little, 2004] Little, J. D. (2004). Models and managers: The concept of a decision calculus. *Management science*, 50(12_supplement):1841–1853. 49

[Loor and Bionaz, 2012] Loor, J. J. and Bionaz, M. (2012). Nutritional genomics: from functional gene networks to feedbunk. 19, 24

[Loor et al., 2011] Loor, J. J., Bionaz, M., and Invernizzi, G. (2011). Systems biology and animal nutrition: insights from the dairy cow during growth and the lactation cycle. *Systems Biology and Livestock Science*. 19, 24

[Ma et al., 2006] Ma, P. C. H., Chan, K. C. C., Yao, X., and Chiu, D. K. Y. (2006). An evolutionary clustering algorithm for gene expression microarray data analysis. *IEEE Transactions on Evolutionary Computation*, 10:296–314. 30, 31

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. of the 5th Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA. 28, 194

[Maddox, 2003] Maddox, B. (2003). The double helix and the'wronged heroine'. *Nature*, 421(6921):407–408. 16

[Madeira and Oliveira, 2004] Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(1):24–45. 32

[Mahanta et al., 2011] Mahanta, P., Ahmed, H., Bhattacharyya, D., and Kalita, J. (2011). Triclustering in gene expression data analysis: A selected survey. In *Emerging Trends and Applications in Computer Science (NCETACS), 2011 2nd National Conference on*, pages 1 –6. 30, 34

[Marakas, 2003] Marakas, G. M. (2003). *Decision support systems in the 21st century*, volume 134. Prentice Hall Upper Saddle River, NJ. 49

[Mark et al., 1994] Mark, S. D., Liu, S. F., Li, J. Y., Gail, M. H., Shen, Q., Dawsey, S. M., Liu, F., Taylor, P. R., Li, B., and Blot, W. J. (1994). The effect of vitamin and mineral supplementation on esophageal cytology: results from the linxian dysplasia trial. *Int J Cancer*, 57(2):162–166. 19, 21

[Maulik and Bandyopadhyay, 2002] Maulik, . U. and Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1650?–1654. 36, 37, 38, 230

[Meilă, 2007] Meilă, M. (2007). Comparing clusterings? an information based distance. *Journal of multivariate analysis*, 98(5):873–895. 38, 39

[Millen et al., 1996] Millen, B. E., Quatromoni, P. A., Gagnon, D. R., Cupples, L. A., Franz, M. M., and D'Agostino, R. B. (1996). Dietary patterns of men and women suggest targets for health promotion: the framingham nutrition studies. *Am J Health Promot*, 11(1):42–52. 18

[Milligan and Cooper, 1985] Milligan, . G. and Cooper, M. (1985). An examination of procedures for determining the number of clusters in a dataset. *Psychometrika*, 50:159?–179. 36, 37, 38, 230

[Mitchell et al., 2011] Mitchell, J. J., Trakadis, Y. J., and Scriver, C. R. (2011). Phenylalanine hydroxylase deficiency. *Genetics in Medicine*, 13(8):697–707. 17

[Möller-Levet et al., 2003] Möller-Levet, C. S., Klawonn, F., Cho, K.-H., Yin, H., and Wolkenhauer, O. (2003). Clustering of unevenly sampled gene expression time-series data. 30, 31

[Moore and Notz, 2006] Moore, D. S. and Notz, W. I. (2006). *Data ethics*, chapter 7. New York: W.H. Freeman., 6th ed. edition. 20

[Mutch et al., 2005] Mutch, D., Wahli, W., and Williamson, G. (2005). Nutrigenomics and nutrigenetics: the emerging faces of nutrition. *The FASEB journal*, 19(12):1602–1616. 19, 22

[National Tech Center, 2013] National Tech Center (2013). National center for technology innovation: Experimental study design. `http://www.nationaltechcenter.org/index.php/products/at-research-matters/experimental-study-design/`. 19

[NuGO, 2015] NuGO (2015). Nugo is an association of universities and research institutes focusing on the joint development of the research area of molecular nutrition, personalised nutrition, nutrigenomics and nutritional systems biology. `http://www.nugo.org/`. Accessed: 2015-07-31. 18

[Nutrigenomics New Zealand, 2015] Nutrigenomics New Zealand (2015). Nutrigenomics new zealand. `http://www.nutrigenomics.org.nz/`. Accessed: 2015-07-31. 18

[Ordovas and Mooser, 2007] Ordovas, J. M. and Mooser, V. (2007). Nutrigenetics and nutrigenomics. *Casopis Lekaru Ceskych*, 146(2):837–839. 5, 19, 22

[Pal and Biswas, 1997] Pal, N. R. and Biswas, J. (1997). Cluster validation using graph theoretic concepts. *Pattern Recognition*, 30(6):847–857. 41

[Panagiotou and Nielsen, 2009] Panagiotou, G. and Nielsen, J. (2009). Nutritional systems biology: definitions and approaches. *Annual Review of Nutrition*, 29(April):329–339. 5, 14

[Pearson, 1901] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572. 46

[Perona et al., 2011] Perona, J. S., Fitó, M., Covas, M.-I., Garcia, M., and Ruiz-Gutierrez, V. (2011). Olive oil phenols modulate the triacylglycerol molecular species of human very low-density lipoprotein. a randomized, crossover, controlled trial. *Metabolism Clinical And Experimental*, 60(6):893–899. 19, 24

[Plackett, 1983] Plackett, R. L. (1983). Karl pearson and the chi-squared test. *International Statistical Review / Revue Internationale de Statistique*, 51(1):59–72. 47

[Poch et al., 2004] Poch, M., Comas, J., Rodriguez-Roda, I., Sànchez-Marrè, M., and Cortés, U. (2004). Designing and building real environmental decision support systems. *Environmental Modelling & Software*, 19(9):857–873. 49

[Power, 2008] Power, D. J. (2008). Decision support systems: a historical overview. In *Handbook on Decision Support Systems 1*, pages 121–140. Springer. 49

[PREDIMED, 2015] PREDIMED (2015). Effects of the mediterranean diet on the primary prevention of cardiovascular diseases. `http://predimed.onmedic.net/`. Accessed: 2015-07-31. 6, 24

[Prelic' et al., 2006] Prelic', A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129. 30, 32

[R Development Core Team, 2012] R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 83, 94, 95

[Ralambondrainy, 1988] Ralambondrainy, H. (1988). *A clustering method for nominal data and mixture ...* H.H.Bock, Elsevier Science Publishers, B.V. (North-Holland). 36

[Ralambondrainy, 1995] Ralambondrainy, H. (1995). *A conceptual version of the K-means algorithm*. Lifetime Learning Publications, Belmont, California. 36

[Robinson, 2010] Robinson, T. R. (07/04/2010). *Genetics For Dummies.* John Wiley & Sons. 16

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65. 41

[Rudolph and Gibert, 2014] Rudolph, A. G. and Gibert, K. (2014). A data mining approach to identify cognitive neurorehabilitation range in traumatic brain injury patients. *Expert Syst. Appl.*, 41:5238–5251. 9

[Salas-Salvadó et al., 2011] Salas-Salvadó, J., Bulló, M., Babio, N., Martínez-González, M. n., Ibarrola-Jurado, N., Basora, J., Estruch, R., Covas, M.-I., Corella, D., Arós, F., and et al. (2011). Reduction in the incidence of type 2 diabetes with the mediterranean diet: results of the predimed-reus nutrition intervention randomized trial. *Diabetes Care*, 34(1):14–9. 19, 24

[Sarstedt and Mooi, 2011] Sarstedt, M. and Mooi, E. (2011). *A Concise Guide to Market research: The process, data, and methods using IBM SPSS statistics.* Springer Verlag. 44

[Schröder et al., 2011] Schröder, H., Fitó, M., Estruch, R., Martínez-González, M. A., Corella, D., Salas-Salvadó, J., Lamuela-Raventós, R., Ros, E., Salaverría, I., Fiol, M., et al. (2011). A short screener is valid for assessing mediterranean diet adherence among older spanish men and women. *The Journal of nutrition*, 141(6):1140–1145. 102

[Segal et al., 2003] Segal, E., Battle, A., and Koller, D. (2003). Decomposing gene expression into cellular processes. In *Proc. Pacific Symp. Biocomputing*, volume 8, pages 89–100. 17

[Sevilla-Villanueva et al., 2013] Sevilla-Villanueva, B., Gibert, K., and Sànchez-Marrè, M. (2013). Post-processing the class panel graphs: Towards understandable patterns from data. In *CCIA*, pages 215–224. 9, 10

[Sevilla-Villanueva et al., 2014] Sevilla-Villanueva, B., Gibert, K., and Sànchez-Marrè, M. (2014). Nutrition pattern assessment through integrative multiview clustering. *submitted to Artificial Intelligence in Medicine.* 10

[Sibson, 1973] Sibson, R. (1973). Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34. 194

[Siponen et al., 2001] Siponen, M., Vesanto, J., Simula, O., and Vasara, P. (2001). An approach to automated interpretation of som. In *Advances in Self-Organising Maps*, pages 89–94. Springer. 44

[Sànchez-Marrè et al., 1999] Sànchez-Marrè, M., Cortés, U., Roda, I. R., and Poch, M. (March 1999). Sustainable case learning for continuous domains. In *Environmental Modelling and Software*, volume 14, pages 349–357. 36

[Sànchez-Marrè et al., 1998] Sànchez-Marrè, M., Roda, I. R., and Comas, Q. (1998). L'eixample distance: a new similarity measure for case retrieval. *1st Catalan Conference on Artificial Intelligence(CCIA'98), ACIA bulletin*, 14-15:246–253. 36

[Swaminathan et al., 2012] Swaminathan, S., Thomas, T., Yusuf, S., and Vaz, M. (2012). Clustering of diet, physical activity and overweight in parents and offspring in south india. *European Journal of Clinical Nutrition*. 29

[Tai et al., 2005] Tai, E. S., Corella, D., Demissie, S., Cupples, L. A., Coltell, O., Schaefer, E. J., Tucker, K. L., and Ordovas, J. M. (2005). Polyunsaturated fatty acids interact with the ppara-l162v polymorphism to affect plasma triglyceride and apolipoprotein c-iii concentrations in the framingham heart study. *The Journal of nutrition*, 135(3):397–403. 19, 22

[Tai and Chiu, 2009] Tai, Y.-M. and Chiu, H.-W. (2009). Comorbidity study of adhd: Applying association rule mining (arm) to national health insurance database of taiwan. *International Journal of Medical Informatics*, 78(12):e75 – e83. <ce:title>Mining of Clinical and Biomedical Text and Data Special Issue</ce:title>. 19, 21

[Talavera and Gaudioso, 2004] Talavera, L. and Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In *Workshop on artificial intelligence in CSCL. 16th European conference on artificial intelligence*, pages 17–23. 44

[Tamayo et al., 1999] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96(6):2907–2912. 30, 31

[Tanay et al., 2002] Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. In *In Proceedings of ISMB 2002*, pages 136–144. 30, 32

[Taylor et al., 1994] Taylor, P. R., Li, B., Dawsey, S. M., Li, J. Y., Yang, C. S., Guo, W., and Blot, W. J. (1994). Prevention of esophageal cancer: the nutrition intervention trials in linxian, china. linxian nutrition intervention trials study group. *Cancer Res*, 54(7 Suppl):–2031. 19, 21

[Tucker et al., 2005] Tucker, K. L., Hannan, M. T., Qiao, N., Jacques, P. F., Selhub, J., Cupples, L. A., and Kiel, D. P. (2005). Low plasma vitamin b12 is associated with lower bmd: The framingham osteoporosis study. *Journal of Bone and Mineral Research*, 20(1):152–158. 19, 22

[van't Veer et al., 2002] van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530– 536. 30

[Venables and Ripley, 2002] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0. 97

[Venter et al., 2001] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., and et al. (2001). The sequence of the human genome. *Science*, 291:1304–1351. 17

[Viguerie et al., 2005] Viguerie, N., Poitou, C., Cancello, R., Stich, V., Clément, K., and Langin, D. (2005). Transcriptomics applied to obesity and caloric restriction. *Biochimie*, 87(1):117 – 123. 19, 25

[Ward Jr, 1963] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244. 28

[WHO, 2015] WHO (2015). World health organization. http://www.who.int/. Accessed: 2015-07-31. 17

[Wickham, 2012] Wickham, H. (2012). *stringr: Make it easier to work with strings.* R package version 0.6.2. 96, 97

[Wilson and Martinez, 1997] Wilson, D. R. and Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of artificial intelligence research*, pages 1–34. 194

[Xu and Wunsch, 2010] Xu, R. and Wunsch, D. C. (2010). Clustering algorithms in biomedical research: a review. *IEEE Rev Biomed Eng*, 3:120–154. 30, 32

[Yang et al., 2003] Yang, J., Wang, H., Wang, W., Yu, P., Ibm, U., Chapel, U., Ibm, H., Watson, T. J., and Watson, T. J. (2003). Enhanced biclustering on expression data. In *Proc. of 3rd IEEE Symposium on BioInformatics and BioEngineering (BIBE'03*, pages 321–327. 30, 33

[Yang et al., 2002] Yang, J., Wang, W., Wang, H., and Yu, P. (2002). Improving performance of bicluster discovery in a large data set. 30, 33

[Yin et al., 2005] Yin, X., Han, J., and Yu, P. S. (2005). Cross-relational clustering with user's guidance. In *ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 344–353. ACM. 29

[Yoo et al., 2012] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., and Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst*, 36(4):2431–2448. 30

[Zhao and Zaki, 2005] Zhao, L. and Zaki, M. J. (2005). tricluster: An effective algorithm for mining coherent clusters in 3d microarray data. In *In Proc. of the 2005 ACM SIGMOD international conference on Management of data*, pages 694–705. ACM Press. 30, 34

[Zimek, 2008] Zimek, A. (2008). *Correlation Clustering*. PhD thesis, Fakultät für Mathematik, Informatik und Statistik der Ludwig–Maximilians–Universität München. 33

[Zou et al., 2002] Zou, X. N., Taylor, P. R., Mark, S. D., Chao, A., Wang, W., Dawsey, S. M., Wu, Y. P., Qiao, Y. L., and Zheng, S. F. (2002). Seasonal variation of food consumption and selected nutrient intake in linxian, a high risk area for esophageal cancer in china. *Int J Vitam Nutr Res*, 72(6):375–382. 19, 21

Glossary

biomarkers Biomarkers (short for biological markers) are biological measures of a biological state. By definition, a biomarker is ”a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention”. Biomarkers are the measures used to perform a clinical assessment such as blood pressure or cholesterol level and are used to monitor and predict health states in individuals or across populations so that appropriate therapeutic intervention can be planned. Biomarkers may be used alone or in combination to assess the health or disease state of an individual.. 13

genome A genome is an organism’s complete set of DNA, including all of its genes. Each genome contains all of the information needed to build and maintain that organism. In humans, a copy of the entire genome -more than 3 billion DNA base pairs- is contained in all cells that have a nucleus.. 3

metabolome The metabolome refers to the complete set of small-molecule chemicals found within a biological sample. The biological sample can be a cell, a cellular organelle, an organ, a tissue, a tissue extract, a biofluid or an entire organism. The small molecule chemicals found in a given metabolome may include both endogenous metabolites that are naturally produced by an organism (such as amino acids, organic acids, nucleic acids, fatty acids, amines, sugars, vitamins, co-factors, pigments, antibiotics, etc.) as well as exogenous chemicals (such as drugs, environmental contaminants, food additives, toxins and other xenobiotics) that are not naturally produced by an organism. The word metabolome appears to be a blending of the words "metabolite" and "chromosome". It was constructed to imply that metabolites are indirectly encoded by genes or act on genes and gene products.. 3

proteome The term proteome refer to the total protein-encoding capability of an organism’s genome. In use, it is also used to refer to the actual total expressed protein complement of an organism or to the identifiable proteins of a single tissue or cell type or subcellular compartment. While the genome is static, the proteome continually changes in response to external and internal events.. 3

systems biology Systems biology is the computational and mathematical modeling of complex biological systems. An emerging engineering approach applied to biomedical and biological scientific research, systems biology is a biology-based inter-disciplinary field of study that focuses on complex interactions within biological systems, using a holistic approach (holism instead of the more traditional reductionism) to biological and biomedical research. . 11

Appendix A

Attribute Dictionary

Attribute Dictionary

id (id) Identifier

grup_ int (dietGroup) Diet intervention group assigned

grup_ dm (dietGroupMed) Diet intervention group assigned

nodo Node

centro Center

paciente (id) Identifier

tipoparticip

seguimiento

dianac (birthDay) INC: Birth Day

mesnac (birthMonth) INC: Birth Month

anyonac (birthYear) INC: Birth Year

sexo (gender) INC: Gender

diaexam0 (inclusionExamDay) INC: Exam Day

mesexam0 (inclusionExamMonth) INC: Exam Month

anyoexam0 (inclusionExamYear) INC: Exam Year

fechaexam0 (inclusionExamDate) INC: Exam Date

proceden INC: Origin

c_ grasan (avoidAnimalFat) INC: Do you usually avoid to eat much animal fat (butter, shortening, bakery ...)? If not, would you be willing to try?

c_ fruta (richFiber) INCLUSION: Do you follow a diet rich in fiber, ie with plenty of fruit, vegetables and legumes? If not, would you be willing to try?

cambdon INCLUSION: Do you think moving to another town in the next few years or have a limitation that prevents or hinders controls and attend scheduled meetings?

prob_ die INCLUSION: Do you informed by health workers suffering from an illness that prevents you follow any particular diet that includes olive oil and / or nuts it has been?

iam_ 0 INCLUSION: Do you informed by health personnel who have ever had a myocardial infarction has been?

angor_ 0 INCLUSION: Do you informed by medical personnel ever had angina has been?

avc_ 0 INCLUSION: Do you informed by health personnel who have ever had a stroke or a stroke has been?

arritmia INCLUSION: Do you informed by medical personnel ever had any heart disease or arrhythmias has been?

tipo_ arr INCLUSION: Diagnosis arrhythmia

claudica INCLUSION: Do you informed by medical personnel ever had has been intermittent claudication?

diabetes INCLUSION: Do you informed by health workers who have had diabetes has been?

ano_ diag INCLUSION: Approximate years of diagnosis of diabetes

hipercol (hyperCholesterolemia) INCLUSION: Do you informed by health workers of having high cholesterol?

tra_ col (medhyperCholesterolemia) INCLUSION: Do you follow any Hypolipidemic agents treatment?

col_ inc (cholesterol) INCLUSION: If yes, annotate Total Col.

hdl_ inc (HDL) INCLUSION: If yes, annotate HDL Col.

idl_ inc (LDL) INCLUSION: If yes, annotate LDL Col.

trigl_ in (tryglicerids) INCLUSION: If yes, annotate Triglycerides

hta0 (hyperTension) INCLUSION: Have you been informed by health workers of having high blood pressure?

trathta0 (medHyperTension) INCLUSION: Do you follow any antihypertensive therapy?

psist0 (sistolicPressure) INCLUSION: If yes, note systolic blood pressure

pdias0 (distolicPressure) INCLUSION: If yes, note diastolic blood pressure

ant_ iam (myocardialInfartion) INCLUSION: Any immediate family (parents, brothers, sons, uncles) have suffered or died from a heart attack or angina at an earlier age to 55 years (men)/65 (women)?

tabaco (tobacco_1) INCLUSION: Do you smoke cigarettes now?

anyos_ tab (yearsSmoking) INCLUSION: If yes, how many years you smoke?

cigarril (numCigarettes_1) INCLUSION: About how many cigarettes do you smoke per day?

puros INCLUSION: About how many cigars do you smoke per day?

pipas INCLUSION: About how many pipes do you smoke per day?

cap_ cam (adaptDiet) INCLUSION: Are you able to change/follow the advised diet by the doctors study?

peso_ inc (weight) INCLUSION: Weight

altura (height) INCLUSION: Height

imc_ inc (BMI) INCLUSION: Body Mass Index

inclus Inclusion: Inclusion

motiv_ ex INCLUSION: Reason for exclusion

excl_ mot INCLUSION: Another reason for exclusion

cont_ ap1 Surname

cont_ nom Name

cont_ ap2 Second surname

con_ tel1 Phone 1

con_ tel2 Phone 2

est_ civi (civilState) GENERAL: Civil Status

escolar (estudies) GENERAL: What is the highest degree of education?

numper (numPerson) GENERAL: Number of people with whom you share home

paisnac (Country) GENERAL: Country (only foreigners)

lugarnac (birthPlace) GENERAL: Birth place

cip CIP

nif DNI

diaexam1 (gralVisitDay) GENERAL: Visit Day

mesexam1 (gralVisitMonth) GENERAL: Visit Month

anyoexam1 (gralVisitYear) GENERAL: Visit Year

fechaexam1 (gralVisitDate) GENERAL: Visit Date

sitlabor (workStatus) GENERAL: What is your current employment status?

pertensa (stressful) GENERAL: Do you consider being a tense and/or aggressive person? (From 0 (relaxed) to 10 (most competitive))

embolia GENERAL: Diagnosed pulmonary embolism?

aneuris GENERAL: Diagnosed aortic aneurysm?

icizq GENERAL: Diagnosed left heart failure?

troboli GENERAL: Diagnosed deep vein thrombosis?

bronqui GENERAL: Diagnosed chronic bronchitis - emphysema?

depre (hasDepression) GENERAL: Diagnosed depression?

catarata GENERAL: Diagnosed Cataracts?

apneas GENERAL: Diagnosed sleep apnea?

cancer (hasCancer) GENERAL: Diagnosed cancer or tumors?

fractura (hasBoneFracture) GENERAL: Diagnosed bone fractures?

demenc GENERAL: Diagnosed dementia?

parckin GENERAL: Diagnosed with Parkinson disease?

retino GENERAL: Diagnosed Retinopathy?

cardio GENERAL: Diagnosed vascular disease?

nefro GENERAL: Diagnosed nephropathy?

fam_col (famColestherol) GENERAL: Any immediate family (parents, siblings, children, etc.) have high cholesterol?

fam_hta (famBloodPressure) GENERAL: Any immediate family (parents, siblings, children, etc.) have high blood pressure?

fam_can (famCancer) GENERAL: Any immediate family (parents, siblings, children, etc.) has or had cancer?

disnea (hasDyspnea_1) GENERAL: Have you noticed in the last year, you too tired or short of breath when you do some exercise (climbing stairs, walking, etc)?

fam_exit (famHeartDiseases) GENERAL: Any immediate family (parents, siblings, children, etc.) died from cardiac causes, or had any heart problems?

fam_avc (cerebrovascularAccident) GENERAL: Any immediate family (parents, siblings, children, etc.) has had a stroke?

trab_cab (famWork) GENERAL: What job do or did the head of household?

trab_pac (work) GENERAL: What job you do or did?

tipo_can (cancerType) GENERAL: Specify the type of cancer or tumor

edad_em GENERAL: Age at diagnosis of pulmonary embolism

edad_ane GENERAL: Age at diagnosis of aortic aneurysm

edad_ic GENERAL: Age at diagnosis of left heart failure

edad_tro GENERAL: Age at diagnosis of deep vein thrombosis

edad_fra (fracturaAge) GENERAL: Age at diagnosis of bone fractures

edad_ret GENERAL: Age at diagnosis of retinopathy

edad_car GENERAL: Age ar diagnosis of vascular disease

edad_nef GENERAL: Age at diagnosis of nephropathy

eda_epoc GENERAL: Age at diagnosis of chronic bronchitis - emphysema

edad_dep (depressionAge) GENERAL: Age at diagnosis of depression

edad_cat GENERAL: Age at diagnosis of cataracts

edad_apn GENERAL: Age at diagnosis of sleep apnea

edad_dem GENERAL: Age at diagnosis of dementia

edad_par GENERAL: Age at diagnosis of disease Parkinson

edad_can (cancerAge) GENERAL: How old were you when your cancer started?

med1 GENERAL: Name medication 1.

mol_bebe GENERAL: Has it ever bothered you people criticizing your drinking?

beb_mal GENERAL: Have you ever felt bad or guilty about your drinking?

beb_meno GENERAL: Have you had the impression that you should drink less?

beb_many GENERAL: Have you ever drank at first thing in the morning to steady your nerves or to get rid of a hangover?

aspirin1 GENERAL: In the past month have you taken aspirin, Adiro or similar?

aines1 GENERAL: In the past month have you taken other drugs to relieve pain or fever?

tranqui1 GENERAL: In the past month have you taken tranquilizers, sedatives, anxiety pills or sleeping pills?

vitamin1 GENERAL: In the past month have you taken vitamin or mineral?

cardiov1 GENERAL: In the past month have you taken heart medication?

hipoten1 GENERAL: In the past month have you taken medication for high blood pressure?

hipocol1 GENERAL: In the past month have you taken cholesterol medication?

insulin1 GENERAL: In the past month have you taken insulin?

ado1 GENERAL: In the past month have you taken medication for diabetes? (Other than insulin)

hormo1 GENERAL: In the past month have you taken hormone therapy? (Only women)

otromed1 GENERAL: In the past month have you taken any other medicines?

med2 GENERAL: Name medication 2.

med3 GENERAL: Name medication 3.

med4 GENERAL: Name medication 4.

med5 GENERAL: Name medication 5.

med6 GENERAL: Name medication 6.

med8 GENERAL: Name medication 8.

med7 GENERAL: Name medication 7.

med1a GENERAL: Medicine 1. morning doses

med1b GENERAL: Medicine 1. noon doses

med1c GENERAL: Medicine 1. night doses

acei_ di1 (lastOliveOilDay) GENERAL: Last Day of intake of virgin olive oil

acei_ me1 (lastOliveOilMonth) GENERAL: Last month of intake of virgin olive oil

vino_ di1 (lastWineDay) GENERAL: Last day of intake of alcohol or wine

vino_ me1 (lastWineMonth) GENERAL: Last month of intake of alcohol or wine

acei_ ho1 (oliveOilToday) GENERAL: Last intake of virgin olive oil today

vino_ ho1 (wineToday) GENERAL: Last intake of alcohol or wine today

edad_ men (menopauseAge) GENERAL: Only women: When your menopause started?

med7a GENERAL: Medicine 7. morning doses

med7b GENERAL: Medicine 7. noon doses

med7c GENERAL: Medicine 7. night doses

med8a GENERAL: Medicine 8. morning doses

med8b GENERAL: Medicine 8. noon doses

med8c GENERAL: Medicine 8. night doses

med2a GENERAL: Medicine 2. morning doses

med2b GENERAL: Medicine 2. noon doses

med2c GENERAL: Medicine 2. night doses

med3a GENERAL: Medicine 3. morning doses

med3b GENERAL: Medicine 3. noon doses

med3c GENERAL: Medicine 3. night doses

med4a GENERAL: Medicine 4. morning doses

med4b GENERAL: Medicine 4. noon doses

med4c GENERAL: Medicine 4. night doses

med5a GENERAL: Medicine 5. morning doses

med5b GENERAL: Medicine 5. noon doses

med5c GENERAL: Medicine 5. night doses

med6a GENERAL: Medicine 6. morning doses

med6b GENERAL: Medicine 6. noon doses

med6c GENERAL: Medicine 6. night doses

acei_ any1 (lastOliveOilYear) GENERAL: Last year of intake of virgin olive oil

vino_ any1 (lastWineYear) GENERAL: Last year of intake of alcohol or wine

altura1 (height_1) GENERAL: Height

peso1 (weight_1) GENERAL: Weight in Q-General

imc1 (CMI_1) GENERAL: Body Mass Index

cint1 (waist_1) GENERAL: Waist

pas_ esis_ 1 GENERAL: left upper extremity (sitting patient) 1st taking.PAS

pad_ esis_ 1 GENERAL: left upper extremity (sitting patient) 1st taking.PAD

fc_ a_ 1 GENERAL: left upper extremity (sitting patient) 1st taking.FC

pas_ esis_ 2 GENERAL: left upper extremity (patient seated) 2 taking.PAS

pad_ esis_ 2 GENERAL: left upper extremity (patient seated) 2 taking.PAD

fc_b_2 GENERAL: left upper extremity (patient seated) 2 taking.FC

pas_esds_1 (systolicPressure_1) GENERAL: right upper extremity (seated patient) 1st taking.PAS

pad_esds_1 (diastolicPressure_1) GENERAL: right upper extremity (sitting patient) 1st taking.PAD

fc_c_2 (heartRate_1) GENERAL: right upper extremity (sitting patient) 1st taking.FC

pas_esds_2 (systolicPressure_2) GENERAL: right upper extremity (seated patient) 2 taking.PAS

pad_esds_2 (diastolicPressure_2) GENERAL: right upper extremity (seated patient) 2 taking.PAD

fc_d_2 (heartRate_2) GENERAL: right upper extremity (seated patient) 2 taking.FC

inciden_2 GENERAL: Incident notes

visita_1_14p P14 VISIT 1

diap14_1 (firstVisitDay) VISIT P14 1: P14 initial day

mesp14_1 (firstVisitMonth) VISIT P14 1: P14 initial month

anop14_1 (firstVisitYear) VISIT P14 1: P14 initial year

p14_1_1 (mainOliveOil) VISIT P14 1: Do you use olive oil as the main cooking fat? INITIAL

p14_2_1 (oliveOil) VISIT P14 1: How much olive oil consume in total per day? (Including that used for frying, eating out, salads, etc) INITIAL

p14_3_1 (vegetables) VISIT P14 1: How many portions of vegetables you eat per day? (side dish or accompaniments count as half portion) 1 portion = 200g INITIAL

p14_4_1 (fruit) VISIT P14 1: How many pieces of fruit (including natural juices) per day? INITIAL

p14_5_1 (redMeat) VISIT P14 1: How many portions of red meat, burgers, hot dogs or sausages consume per day? 1 portion = INITIAL 100-150g

p14_6_1 (butter) VISIT P14 1: How many portions of butter, margarine or cream per day? INITIAL portion = 12g

p14_7_1 (gasDrinks) VISIT P14 1: How many (sugared) soft drinks (sodas, colas, tonics, bitter) per day? INITIAL

p14_8_1 (wine) VISIT P14 1: Do you drink wine? How much consumes per week? INITIAL

p14_9_1 (legume) VISIT P14 1: How many portions of legumes consumes per week? 1 bowl or portion = 150g INITIAL

p14_10_1 (fish) VISIT P14 1: How many portions of fish/seafood consumes per week? 1 portion = 100-150g of fish or 4-5 pieces or 200gr of seafood INITIAL

p14_11_1 (commercialBakery) VISIT P14 1: How often consume (not homemade) Commercial bakery like cookies, puddings, sweets or cakes per week? INITIAL

p14_12_1 (nuts) VISIT P14 1: How often consume nuts per week? INITIAL 1 portion = 30g

p14_13_1 (whiteMeat) VISIT P14 1: Do you preferably eat meat of chicken, rabbit or turkey instead of beef, pork, burgers and sausages? Chicken: 1 piece or portion 100-150g INITIAL

p14_14_1 (sauce) VISIT P14 1: How many times per week consume cooked vegetables, pasta, rice or other dishes dressed with homemade tomato sauce (garlic, onion or leek made simmered with olive oil)? INITIAL

visita_1_af

dia_af_1 (firstAfVisitDay) AF VISIT 1: Exam day

mes_af_1 (firstAfVisitMonth) AF VISIT 1: Exam month

ano_af_1 (firstAfVisitYear) AF VISIT 1: Exam year

ge2_3_5s_1 (lightWeek) AF VISIT 1: light physical activity (last week) KCAL/day

ge4_5_5s_1 (moderateWeek) AF VISIT 1: moderate physical activity (last week) KCAL/day

ge6_12s_1 (intenseWeek) AF VISIT 1: intense physical activity (last week) KCAL/day

gehos_1 (homeWorkWeek) AF VISIT 1: Physical Activity home (last week) KCAL/day

getots_1 (totalWeek) AF VISIT 1: Total physical activity (last week) KCAL/day

ge2_3_5a_1 (lightYear) AF VISIT 1: light physical activity (last year) KCAL/day

ge4_5_5a_1 (moderateYear) AF VISIT 1: moderate physical activity (last year) KCAL/day

ge6_12a_1 (intenseYear) AF VISIT 1: Intense physical activity (last year) KCAL/day

gehoa_1 (homeWorkYear) AF VISIT 1: Physical Activity home (last year) KCAL/day

getota_1 (totalYear) AF VISIT 1: Total physical activity (last year) KCAL/day

seg Follow (in months)

est_civ2 (civilState_2) FOLLOW VISIT 2: Civil status 3-Month

n_perho (numPerson) FOLLOW VISIT 2: Number of people with whom they share home 3-month

cam_est (changeCivilState) FOLLOW VISIT 2: Has your civil status changed since your last visit? 3 MONTHS

cam_pho (changeNumPerson) FOLLOW VISIT 2: Have you changed the number of people with whom they share the home since the last visit? 3 MONTHS

n_trab_c FOLLOW VISIT 2: What job has the head of household? 3 MONTHS

n_treball (work_2) FOLLOW VISIT 2: What concrete work do or did? 3 MONTHS

sit_labo (workStatus_2) FOLLOW VISIT 2: What is your current employment status? 3 MONTHS

cam_slab (changeWorkStatus) FOLLOW VISIT 2: Have you changed your employment status since the last visit? 3 MONTHS

diaexam2 (secondVisitDay) FOLLOW VISIT 2: Examination day 3-month

mesexam2 (secondVisitMonth) FOLLOW VISIT 2: Examination month 3-month

anyoexam2 (secondVisitYear) FOLLOW VISIT 2: Examination Year 3-month

fechaexam2 (secondVisitDate) FOLLOW VISIT 2 Examination date 3-month

cigarr2 (cigarettes) FOLLOW VISIT 2: How many cigarettes do you smoke per day? 3 MONTHS

puros2 (cigar) FOLLOW VISIT 2: How many cigars do you smoke per day? 3 MONTHS

pipas2 (pipe) FOLLOW VISIT 2: How many pipes do you smoke per day? 3 MONTHS

cam_tab (change _tobacco) FOLLOW VISIT 2: Have you changed your smoking habits in the last 6 months?

iam2 FOLLOW VISIT 2: Have you been informed by health personnel of having a heart attack in the past year? 3 MONTHS

angor2 FOLLOW VISIT 2: Have you been told by medical personnel of having angina in the last year? 3 MONTHS

avc2 FOLLOW VISIT 2: Have you been informed by health personnel of suffering a stroke in the past year has been? 3 MONTHS

diabete2 FOLLOW VISIT 2: Have you been diagnosed in the last year of diabetes? 3 MONTHS

hipecol2 FOLLOW VISIT 2: Have you been informed by health personnel of having high cholesterol? 3 MONTHS

hta2 (hypertension) FOLLOW VISIT 2: Have you been informed by health personnel of having high blood pressure? 3 MONTHS

angiop2 FOLLOW VISIT 2: Have you been informed by health personnel of having a coronary angioplasty with or without stenting or Coron bypass surgery

paro2 FOLLOW VISIT 2: Have you been informed by health personal of suffering a heart attack from which you have recovered in the last year? 3 MONTHS

aneuri2 FOLLOW VISIT 2: Have you been diagnosed in the last year of aortic aneurysm? 3 MONTHS

disnea2 (dyspnoea) FOLLOW VISIT 2: Have you noticed in the last year that you are too tired or fell a lack of air during some exercise (climbing stairs, walking, etc)? 3 MONTHS

claudi2 FOLLOW VISIT 2: In the last year, have been diagnosed with a circulatory disorder in the legs? 3 MONTHS

iqvasc2 FOLLOW VISIT 2: Has this circulatory disorder been the cause of an intervention or amputation of part of a limb? 3 MONTHS

tabaco2 (tobacco) FOLLOW VISIT 2: Do you currently smoke cigarettes? 3 MONTHS

arritm2 FOLLOW VISIT 2: Do you informed by health personnel of suffering an arrhythmia in the last year? 3 MONTHS

acei_di2 (lastOliveOilDay) FOLLOW VISIT 2: Last Day intake of virgin olive oil 3-month

acei_me2 (lastOliveOilMonth) FOLLOW VISIT 2: Last month intake of virgin olive oil 3-month

acei_any2 (lastOliveOilYear) FOLLOW VISIT 2: Last year intake of virgin olive oil 3-month

vino_di2 (lastWineDay) FOLLOW VISIT 2: Last day intake of alcohol or wine 3-month

vino_me2 (lastWineMonth) FOLLOW VISIT 2: Last month intake of alcohol or wine 3-month

vino_any2 (lastWineYear) FOLLOW VISIT 2: Last year intake of alcohol or wine 3-month

acei_ho2 (oliveOilToday) FOLLOW VISIT 2: Today, last intake of virgin olive oil 3-month

vino_ho2 (wineToday) FOLLOW VISIT 2: Today, last intake of alcohol or wine 3-month

nefrodb2 FOLLOW VISIT 2: In diabetic patients, have you been diagnosed in the last year some of the following complications: Renal affectation? 3-month

dialis2 FOLLOW VISIT 2: In case of renal affectation, the deterioration of renal function has caused entering a dialysis program? 3 MONTHS

retinop2 FOLLOW VISIT 2: Affectation of the retina from diabetes (diabetic Retinopathy) that triggered a laser treatment 3-month

catarat2 FOLLOW VISIT 2: Have you been diagnosed in the last year of cataracts? 3 MONTHS

intquir2 FOLLOW VISIT 2: Have you been surgically intervened this past year? 3 MONTHS

tipoint2 FOLLOW VISIT 2: surgery type 3-MONTHS

enferm2 FOLLOW VISIT 2: Have you developed in the last year some kind of disease that you have not previously diagnosed (if yes indicate which) 3-MONTHS?

tipoenf2 FOLLOW VISIT 2: disease type 3-month

cam_med (change_medication) FOLLOW VISIT 2: has changed the type of medication in the past year? 3 MONTHS

aspirin2 FOLLOW VISIT 2: During the past month have you taken aspirin, Adiro or similar? 3 MONTHS

aines2 FOLLOW VISIT 2: During the past month have you taken other drugs to relieve pain or fever? 3 MONTHS

tranqu2 FOLLOW VISIT 2: During the past month have you taken tranquilizers, sedatives, anxiety pills or sleeping pills? 3 MONTHS

vitamin2 FOLLOW VISIT 2: During the past month have you taken vitamin or mineral? 3 MONTHS

cardiov2 FOLLOW VISIT 2: During the past month have you taken heart medication? 3 MONTHS

hipoten2 FOLLOW VISIT 2: During the past month have you taken medication for high blood pressure? 3 MONTHS

hipocol2 FOLLOW VISIT 2: During the past month have you taken cholesterol medication? 3 MONTHS

insulin2 FOLLOW VISIT 2: During the past month have you taken insulin? 3 MONTHS

ado2 FOLLOW VISIT 2: During the past month have you taken medication for diabetes? (Other than insulin) 3 MONTHS

hormo2 FOLLOW VISIT 2: During the past month have you taken hormone therapy? (Women only) 3 MONTHS

otromed2 FOLLOW VISIT 2: During the past month have you taken any other medicines? 3 MONTHS

med21 FOLLOW VISIT 2: Drug Name 1. 3-month

med22 FOLLOW VISIT 2: Drug Name 2. 3-month

med23 FOLLOW VISIT 2: Drug Name 3. 3-month

med24 FOLLOW VISIT 2: Drug Name 4. 3-month

med25 FOLLOW VISIT 2: Drug Name 5. 3-month

med26 FOLLOW VISIT 2: Drug Name 6. 3-month

med28 FOLLOW VISIT 2: Drug Name 7. 3-month

med27 FOLLOW VISIT 2: Drug Name 8. 3-month

med21a FOLLOW VISIT 2: Drung 1. morning doses 3-month

med21b FOLLOW VISIT 2: Drung 1. noon doses 3-month

med21c FOLLOW VISIT 2: Drung 1. night doses 3-month

med22a FOLLOW VISIT 2: Drung 2. morning doses 3-month

med22b FOLLOW VISIT 2: Drung 2. noon doses 3-month

med22c FOLLOW VISIT 2: Drung 2. night doses 3-month

med23a FOLLOW VISIT 2: Drung 3. morning doses 3-month

med23b FOLLOW VISIT 2: Drung 3. noon doses 3-month

med23c FOLLOW VISIT 2: Drung 3. night doses 3-month

med24a FOLLOW VISIT 2: Drung 4. morning doses 3-month

med24b FOLLOW VISIT 2: Drung 4. noon doses 3-month

med24c FOLLOW VISIT 2: Drung 4. night doses 3-month

med25a FOLLOW VISIT 2: Drung 5. morning doses 3-month

med25b FOLLOW VISIT 2: Drung 5. noon doses 3-month

med25c FOLLOW VISIT 2: Drung 5. night doses 3-month

med26a FOLLOW VISIT 2: Drung 6. morning doses 3-month

med26b FOLLOW VISIT 2: Drung 6. noon doses 3-month

med26c FOLLOW VISIT 2: Drung 6. night doses 3-month

med27a FOLLOW VISIT 2: Drung 7. morning doses 3-month

med27b FOLLOW VISIT 2: Drung 7. noon doses 3-month

med27c FOLLOW VISIT 2: Drung 7. night doses 3-month

med28a FOLLOW VISIT 2: Drung 8. morning doses 3-month

med28b FOLLOW VISIT 2: Drung 8. noon doses 3-month

med28c FOLLOW VISIT 2: Drung 8. night doses 3-month

altura2 (height_2) FOLLOW VISIT 2: Height 3-MONTHS

peso2 (weight_2) FOLLOW VISIT 2: Weight 3-MONTHS

imc2 (BMI_2) FOLLOW VISIT 2: Body Mass Index 3-month

cintura2 (waist:2) FOLLOW VISIT 2: Waist 3-month

inciden2 FOLLOW VISIT 2: Incident notes 3-month

grupo (dietGroup) Intervention Groups

oxldl0 (OxidizedLDL_1) Oxidized LDL Baseline (U/l)

oxldl3 (OxidizedLDL_2) Oxidized LDL 3 months (U/l)

hdl0 (HDL_1) Baseline HDL (mg/dl)

hdl3 (HDL_2) HDL 3 months (mg/dl)

gluc0 (glucose_1) Baseline glucose (mg/dl)

gluc3 (glucose_2) 3 months glucose (mg/dl)

cholest0 (cholesterol_1) Baseline cholesterol (mg/dl)

cholest1 (cholesterol_2) 3 months Cholesterol (mg/dl)

pcr0 (CReactiveProtein_1) C-Reactive Protein at baseline

pcr3 (CReactiveProtein_2) C-Reactive Protein at 3 months

tryg0 (triglycerides_1) Baseline Triglycerides (mg/dl)

tryg3 (triglycerides_2) 3 months Triglycerides (mg/dl)

tyru0 (Tyrosol_1) Tyrosol in urine at baseline

tyru3 (Tyrosol_2) Tyrosol in urine at 3 months

ohtyru0 (OHTyrosol_1) OH-tyrosol in urine at baseline

ohtyru3 (OHTyrosol_2) OH-tyrosol in urine at 3 months

mohtyu0 (MOH_tyrosol_1) MOH-tyrosol in urine at baseline

mohtyu3 (MOH_tyrosol_2) MOH-tyrosol in urine at 3 months

il_6_0 (IL6_1) Concentration of IL-6 in plasma (pg/ml) at baseline Measured by Multiplex kit-Flow Cytometry

il_6_3 (IL6_2) Concentration of IL-6 in plasma (pg/ml) at 3 months Measured by Multiplex kit-Flow Cytometry

il_8_0 (IL8_1) Concentration of IL8 in plasma (pg/ml) at baseline Multiplex kit Measured by Flow Cytometry

il_8_3 (IL8_2) Concentration of IL8 in plasma (pg/ml) at 3 months Measured by Multiplex kit-Flow Cytometry

il_10_0 (IL10_1) Concentration of IL10 in plasma (pg/ml) at baseline Measured by Multiplex kit-Flow Cytometry

il_10_3 (IL20_2) Concentration of IL10 in plasma (pg/ml) at 3 months Measured by Multiplex kit-Flow Cytometry

mcp_1_0 (MCP1_1) Concentration of MCP1 in plasma (pg/ml) at baseline Measured by Multiplex kit-Flow Cytometry

mcp_1_3 (MCP2_2) Concentration of MCP1 in plasma (pg/ml) at 3 months Measured by Multiplex kit-Flow Cytometry

t_pa_0 (tPA_1) Concentration of tPA in plasma (pg/ml) at baseline Multiplex kit Measured by Flow Cytometry

t_pa_3 (tPA_2) Concentration of tPA in plasma (pg/ml) at 3 months Measured by Multiplex kit-Flow Cytometry

ifn_g_0 (IFNg_1) Concentration of IFNg in plasma (pg/ml) at baseline Measured by Multiplex kit-Flow Cytometry

ifn_g_3 (IFNg_2) Concentration of IFNg in plasma (pg/ml) at 3 months Measured by Multiplex kit-Flow Cytometry

VAR00002 (dietGroup) GENERAL: Group assigned

s_cd40l_0 (sCD40L_1) Concentration of sCD40L in plasma (pg/ml) before treatment (baseline) Measured by Multiplex kit-Flow Cytometry

s_cd40l_3 (sCD40L_2) Concentration of sCD40L in plasma (pg/ml) after treatment (3 months) Measured by Multiplex kit-Flow Cytometry

tnf_a_0 (TNFa_1) In plasma concentration of TNF (pg/ml) at baseline Measured by Multiplex kit-Flow Cytometry

tnf_a_3 (TNFa_2) In plasma concentration of TNF (pg/ml) at 3 months, Multiplex kit Measured by Flow Cytometry

s_vcam_1_0 (sVCAM1_1) s-vcam Result in ng/ml

s_vcam_1_3 (sVCAM_2) s-vcam Result in ng/ml 3-month

s_p_selectin_0 (sPselectin_1) Concentration of sPselectin in plasma (ng/mL) at baseline, Multiplex kit Measured by Flow Cytometry

s_p_selectin_3 (sPselectin_2) Concentration of sPselectin in plasma (ng/mL) at 3 months, Multiplex kit Measured by Flow Cytometry

oxo_gg_0 (8_OXOgG_1) 8-oxo-gG values at baseline (nmol of 8-oxo-dG/mmol urinary creatinine)

oxo_gg_3 (8_OXOgG_2) 8-oxo-gG at 3weeks values (nmol of 8-oxo-dG/mmol urinary creatinine)

isoprostanos_0 (F2α-Isoprostanes_1) F2α-Isoprostanes values at baseline (pg/mmol creatinine in urine)

isoprostanos_3 (F2α-Isoprostanes_2) F2α-Isoprostanes values at 3 weeks (pg/mmol creatinine in urine)

d_ldlox LDLox difference after - before LDLox

d_hdl difference HDL1-HDL0

d_gluc difference Gluc1-Gluc0

d_cholest difference Cholest1-Cholest0

d_tryg difference Tryg1-Tryg0

d_pes difference PES2-Pes1

d_imc difference BMI1 - BMI0

d_ts (diffSystolicPressure) difference TS0-TS1

d_td (diffDiastolicPressure) difference TD1-TD0

d_pols (diffHeartRate) difference HeartRate1-HeartRate0

d_isoprostanos difference isoprostanes after - before isoprostanes

d_oxogg difference oxogGafter - oxogG before

d_scd40l difference sCD40Lafter-sCD40Lbefore

d_ifng difference IFNgafter-IFNgbefore

d_tpa difference tPAafter-tPAbefore

d_mcp difference MCPafter-MCPbefore

d_il10 difference IL10after-IL10before

d_il8 difference IL8after-IL8before

d_il6 difference IL6after-IL6before

d_spselectin difference sPselectin after - before sPselectin

d_svcam1 difference sVCAM1after- sVCAM1 before

d_tnfa difference TNFa after - before TNFa

abca1antes ABCA1 expression in RQ measured before treatment

abca1desp RQ measured in ABCA1 expression after treatment (3weeks)

abcg1antes ABCG1 expression in RQ measured before treatment (baseline)

abcg1desp ABCG1 measured expression after treatment in RQ (3weeks)

adam17ante ADAM17 expression in RQ measured before treatment (baseline)

adam17desp ADAM17 expression in RQ measured after the treatment (3weeks)

adamts1antes ADAMTS1expression measured tratamienot in RQ before (baseline)

adamts1desp ADAMTS1 measured expression after treatment in RQ (3weeks)

adrb2antes ADRB2 expression in RQ measured before treatment(baseline)

adrb2desp ADRB2 measured expression after treatment in RQ (3weeks)

aldh1a1antes ALDH1A1 expression measured in RQ before treatment

aldh1a1despues Measeured ALDH1A1 expression after treatment in RQ (3weeks)

anxa1antes ANXA1 measured expression in RQ before treatment (baseline)

anxa1despues ANXA1 measured expression after treatment in RQ (3weeks)

arhgap15antes ARHGAP15 measured expression in RQ before treatment (baseline)

arhgap15despues ARHGAP15 measured expression after treatment in RQ (3weeks)

arhgap19antes ARHGAP19 measured expression in RQ before treatment (baseline)

arhgap19despues ARHGAP19 measured expression after treatment in RQ (3weeks)

arhgef6antes ARHGEF6 measured expression in RQ before treatment (baseline)

arhgef6despues ARHGEF6 measured expression after treatment in RQ (3weeks)

ccng1antes CCNG1 measured expression in RQ before treatment (baseline)

ccng1despues CCNG1 measured expression after treatment in RQ (3weeks)

cd36antes CD36 expression in RQ measured before treatment (baseline)

cd36despues RQ measured in CD36 expression after treatment (3weeks)

cetpantes CETP expression in RQ measured before treatment (baseline)

cetpdespues RQ measured in CETP expression after treatment (3weeks)

chukantes CHUK measured expression in RQ before treatment (baseline)

chukdespues CHUK measured expression after treatment in RQ (3weeks)

dclre1cantes DCLRE1C measured expression in RQ before treatment (baseline)

dclre1cdespues DCLRE1C measured expression after treatment in RQ (3weeks)

ercc5antes ERCC5 measured expression in RQ before treatment (baseline)

ercc5despues ERCC5 measured expression after treatment in RQ (3weeks)

ifna1antes IFNA1 measured expression in RQ before treatment (baseline)

ifna1despues IFNA1 measured expression after treatment in RQ (3weeks)

ifngantes IFNG expression measured in RQ before treatment (baseline)

ifngdespues Measeured IFNG expression after treatment in RQ (3weeks)

il10antes IL10 expression in RQ measured before treatment (baseline)

il10despues Measeured IL10 expression after treatment in RQ (3weeks)

il6antes IL6 expression in RQ measured before treatment (baseline)

il6despues Measeured IL6 expression after treatment in RQ (3weeks)

il7rantes IL7R measured expression in RQ before treatment (baseline)

il7rdespues IL7R measured expression after treatment in RQ (3weeks)

liasantes LIAS measured expression in RQ before treatment (baseline)

liasdespues LIAS measured expression after treatment in RQ (3weeks)

mpoantes MPO expression in RQ measured before treatment (baseline)

mpodepues Measeured MPO expression after treatment in RQ (3weeks)

msr1antes MSR1 measured expression in RQ before treatment (baseline)

msr1despues MSR1 measured expression after treatment in RQ (3weeks)

nfkb2antes NFKB2 measured expression in RQ before treatment (baseline)

nfkb2despues NFKB2 measured expression after treatment in RQ (3weeks)

nos2aantes NOS2A measured expression in RQ before treatment (baseline)

nos2adespues NOS2A measured expression after treatment in RQ (3weeks)

nox1ante NOX1 measured expression in RQ before treatment (baseline)

nox1despues NOX1 measured expression after treatment in RQ (3weeks)

nr1h2antes NR1H2 measured expression in RQ before treatment (baseline)

nr1h2despues NR1H2 measured expression after treatment in RQ (3weeks)

nr1h3antes NR1H3 measured expression in RQ before treatment (baseline)

nr1h3despues NR1H3 measured expression after treatment in RQ (3weeks)

ogtantes RQ measured in OGT expression before treatment (baseline)

ogtdespues RQ measured in OGT expression after treatment (3weeks)

olr1antes OLR1 measured expression in RQ before treatment (baseline)

olr1despues OLR1 measured expression after treatment in RQ (3weeks)

osbpantes OSBP measured expression in RQ before treatment (baseline)

osbpdespues OSBP measured expression after treatment in RQ (3weeks)

pla2g4bantes PLA2G4B measured expression in RQ before treatment (baseline)

pla2g4bdespues PLA2G4B measured expression after treatment in RQ (3weeks)

polkantes POLK measured expression in RQ before treatment (baseline)

polkdespues POLK measured expression after treatment in RQ (3weeks)

pparaantes PPARA measured expression in RQ before treatment (baseline)

pparadespues PPARA measured expression after treatment in RQ (3weeks)

pparbpantes PPARBP measured expression in RQ before treatment (baseline)

pparbpdespues PPARBP measured expression after treatment in RQ (3weeks)

ppardantes PPARd measured expression in RQ before treatment (baseline)

pparddespues PPARd measured expression after treatment in RQ (3weeks)

ppargantes PPARG measured expression in RQ before treatment (baseline)

ppargdespues PPARG measured expression after treatment in RQ (3weeks)

ptgs1antes Ptgs1 measured expression in RQ before treatment (baseline)

ptgs1despues Ptgs1 measured expression after treatment in RQ (3weeks)

ptgs2antes PTGS2 measured expression in RQ before treatment (baseline)

ptgs2despues PTGS2 measured expression after treatment in RQ (3weeks)

rgs2antes RGS2 measured expression in RQ before treatment (baseline)

rgs2despues RGS2 measured expression after treatment in RQ (3weeks)

scarb1antes SCARB1 measured expression in RQ before treatment (baseline)

scarb1despues SCARB1 measured expression after treatment in RQ (3weeks)

tnfsf10antes TNFSF10 measured expression in RQ before treatment (baseline)

tnfsf10despus TNFSF10 measured expression after treatment in RQ (3weeks)

tnfsf12_ tnfsf13antes TNFSF12_TNFSF13 measured expression in RQ before treatment (baseline)

tnfsf12_ tnfsf13despues TNFSF12_TNFSF13 measured expression after treatment in RQ (3weeks)

tp53antes TP53 expression in RQ measured before treatment (baseline)

tp53despues Measeured TP53 expression after treatment in RQ (3weeks)

usp48antes USP48 expression in RQ measured before treatment (baseline)

usp48despues Measeured USP48 expression after treatment in RQ (3weeks)

xrcc5antes XRCC5 measured expression in RQ before treatment (baseline)

xrcc5despues XRCC5 measured expression after treatment in RQ (3weeks)

ratio_ rq_ abca1 RATIO_RQ_ABCA1 AFTER/BEFORE

ratio_ rq_ abcg1 RATIO_RQ_ABCG1 after/before

ratio_ rq_ adam17 RATIO_RQ_ADAM17 after/before

ratio_ rq_ adamts1 RATIO_RQ_ADAMTS1 after/before

ratio_ rq_ adrb2 RATIO_RQ_ADRB2 after/before

ratio_ rq_ aldh1a1 RATIO_RQ_ALDH1A1 after/before

ratio_ rq_ anxa1 RATIO_RQ_ANXA1 after/before

ratio_ rq_ arhgap15 RATIO_RQ_ARHGAP15 after/before

ratio_ rq_ arhgap19 RATIO_RQ_ARHGAP19 after/before

ratio_ rq_ arhgef6 RATIO_RQ_ARHGEF6 after/before

ratio_ rq_ ccng1 RATIO_RQ_CCNG1 after/before

ratio_ rq_ cd36 RATIO_RQ_CD36 after/before

ratio_ rq_ cetp RATIO_RQ_CETP after/before

ratio_ rq_ chuk RATIO_RQ_CHUK after/before

ratio_ rq_ dclre1c RATIO_RQ_DCLRE1C after/before

ratio_ rq_ ercc5 RATIO_RQ_ERCC5 after/before

ratio_ rq_ ifna1 RATIO_RQ_IFNA1 after/before

ratio_ rq_ ifng RATIO_RQ_IFNG after/before

ratio_ rq_ il10 RATIO_RQ_IL10 after/before

ratio_ rq_ il6 RATIO_RQ_IL6 after/before

ratio_ rq_ il7r RATIO_RQ_IL7R after/before

ratio_ rq_ lias RATIO_RQ_LIAS after/before

ratio_ rq_ mpo RATIO_RQ_MPO after/before

ratio_ rq_ msr1 RATIO_RQ_MSR1 after/before

ratio_ rq_ nfkb2 RATIO_RQ_NFKB2 after/before

ratio_ rq_ nos2a RATIO_RQ_NOS2A after/before

ratio_ rq_ nox1 RATIO_RQ_NOX1 after/ants

ratio_ rq_ nr1h2 RATIO_RQ_NR1H2 after/before

ratio_ rq_ nr1h3 RATIO_RQ_NR1H3 after/before

ratio_ rq_ ogt RATIO_RQ_OGT after/before

ratio_rq_olr1 RATIO_RQ_OLR1 after/before

ratio_rq_osbp RATIO_RQ_OSBP after/before

ratio_rq_pla2g4b RATIO_RQ_PLA2G4B after/before

ratio_rq_polk RATIO_RQ_POLK after/before

ratio_rq_ppara RATIO_RQ_PPARA after/before

ratio_rq_pparbp RATIO_RQ_PPARBP after/before

ratio_rq_ppard RATIO_RQ_PPARD after/before

ratio_rq_pparg RATIO_RQ_PPARG after/before

ratio_rq_ptgs1 RATIO_RQ_PTGS1 after/before

ratio_rq_ptgs2 RATIO_RQ_PTGS2 after/before

ratio_rq_rgs2 RATIO_RQ_RGS2 after/before

ratio_rq_scarb1 RATIO_RQ_SCARB1 after/before

ratio_rq_tnfsf10 RATIO_RQ_TNFSF10 after/before

ratio_rq_tnfsf12_13 RATIO_RQ_TNFSF12_13 after/before

ratio_rq_tp53 RATIO_RQ_TP53 after/before

ratio_rq_usp48 RATIO_RQ_USP48 after/before

ratio_rq_xrcc5 RATIO_RQ_XRCC5 after/before

log2ratiorq_abca1 log2ratioRQ_ABCA1

log2ratiorq_abcg1 log2ratioRQ_ABCG1

log2ratiorq_adam17 log2ratioRQ_ADAM17

log2ratiorq_adamts1 log2ratioRQ_ADAMTS1

log2ratiorq_adrb2 log2ratioRQ_ADRB2

log2ratiorq_aldh1a1 log2ratioRQ_ALDH1A1

log2ratiorq_anxa1 log2ratioRQ_ANXA1

log2ratiorq_arhgap15 log2ratioRQ_ARHGAP15

log2ratiorq_arhgap19 log2ratioRQ_ARHGAP19

log2ratiorq_arhgef6 log2ratioRQ_ARHGEF6

log2ratiorq_ccng1 log2ratioRQ_CCNG1

log2ratiorq_cd36 log2ratioRQ_CD36

log2ratiorq_cetp log2ratioRQ_CETP

log2ratiorq_chuk log2ratioRQ_CHUK

log2ratiorq_dclre1c log2ratioRQ_DCLRE1C

log2ratiorq_ercc5 log2ratioRQ_ERCC5

log2ratiorq_ifna1 log2ratioRQ_IFNA1

log2ratiorq_ifng log2ratioRQ_IFNG

log2ratiorq_il10 log2ratioRQ_IL10

log2ratiorq_il6 log2ratioRQ_IL6

log2ratiorq_il7r log2ratioRQ_IL7R

log2ratiorq_lias log2ratioRQ_LIAS

log2ratiorq_mpo log2ratioRQ_MPO

log2ratiorq_msr1 log2ratioRQ_MSR1

log2ratiorq_ nfkb2 log2ratioRQ_NFKB2

log2ratiorq_ nos2a log2ratioRQ_NOS2A

log2ratiorq_ nox1 log2ratioRQ_NOX1

log2ratiorq_ nr1h2 log2ratioRQ_NR1H2

log2ratiorq_ nr1h3 log2ratioRQ_NR1H3

log2ratiorq_ ogt log2ratioRQ_OGT

log2ratiorq_ olr1 log2ratioRQ_OLR1

log2ratiorq_ osbp log2ratioRQ_OSBP

log2ratiorq_ pla2g4b log2ratioRQ_PLA2G4B

log2ratiorq_ polk log2ratioRQ_POLK

log2ratiorq_ ppara log2ratioRQ_PPARA

log2ratiorq_ pparbp log2ratioRQ_PPARBP

log2ratiorq_ ppard log2ratioRQ_PPARD

log2ratiorq_ pparg log2ratioRQ_PPARG

log2ratiorq_ ptgs1 log2ratioRQ_PTGS1

log2ratiorq_ ptgs2 log2ratioRQ_PTGS2

log2ratiorq_ rgs2 log2ratioRQ_RGS2

log2ratiorq_ scarb1 log2ratioRQ_SCARB1

log2ratiorq_ tnfsf10 log2ratioRQ_TNFSF10

log2ratiorq_ tnfsf12_ 13 log2ratioRQ_TNFSF12_13

log2ratiorq_ tp53 log2ratioRQ_TP53

log2ratiorq_ usp48 log2ratioRQ_USP48

log2ratiorq_ xrcc5 log2ratioRQ_XRCC5

visita_ 2_ 14p VISIT P14 2

diap14_ 2 (secondp14VisitDay) P14 VISIT 2: P14 day 3-month

mesp14_ 2 (secondp14Visitmonth) P14 VISIT 2: P14 month 3-month

anop14_ 2 (secondp14VisitYear) P14 VISIT 2: P14 year 3-month

p14_ 1_ 2 (mainOliveOil) P14 VISIT 2: Do you use olive oil as the main cooking fat? 3 MONTHS

p14_ 2_ 2 (oliveOil) P14 VISIT 2: How much olive oil consumed in total per day? (Including that used for frying, eating out, salads, etc.) 3 MONTHS

p14_ 3_ 2 (vegetables) P14 VISIT 2: How many portions of vegetables per day? (side dish or accompaniments count as half portion) 1 portion = 200g 3-month

p14_ 4_ 2 (fruit) P14 VISIT 2: How many pieces of fruit (including natural juices) per day? 3 MONTHS

p14_ 5_ 2 (redMeat) P14 VISIT 2: How many portions of red meat, burgers, hot dogs or sausages consume per day? 1 portion = 100-150g 3-month

p14_ 6_ 2 (butter) P14 VISIT 2: How many portions of butter, margarine or cream per day? portion = 12g 3-MONTHS

p14_ 7_ 2 (gasDrinks) P14 VISIT 2: How many (sugared) soft drinks (sodas, colas, tonics, bitter) per day? 3 MONTHS

p14_ 8_ 2 (wine) P14 VISIT 2: Do you drink wine? How much consumes per week? 3 MONTHS

p14_ 9_ 2 (legume) P14 VISIT 2: How many portions of legumes consumed per week? 1 bowl or portion = 150g 3-month

p14_10_2 (fish) P14 VISIT 2: How many portions of fish/seafood consumed per week? 1 portion = 100-150g fish or or 4 - 5 pieces or 200g of seafood 3-month

p14_11_2 (commercialBakery) P14 VISIT 2: How often consumed (not homemade) Commercial bakery like cookies, puddings, sweets or cakes a week? 3 MONTHS

p14_12_2 (nuts) P14 VISIT 2: How often consume nuts per week? 1 portion = 30g 3-month

p14_13_2 (whiteMeat) P14 VISIT 2: Do you preferably eat meat of chicken, rabbit or turkey instead of beef, pork, burgers and sausages? Chicken meat: 1 portion of 100-150g 3-Month

p14_14_2 (sauce) P14 VISIT 2: How many times a week consume cooked vegetables, pasta, rice or other dishes dressed with homemade tomato sauce (garlic, onion or leek made simmered with olive oil)? 3 MONTHS

visita_2_af

dia_af_2 (secondAfVisitDay) AF VISIT 2: Examination day

mes_af_2 (secondAfVisitMonth) AF VISIT 2: Examination month

ano_af_2 (secondAfVisityear) AF VISIT 2: Examination year

ge2_3_5s_2 (lightWeek) AF VISIT 2: light physical activity (last week) KCAL/day

ge4_5_5s_2 (moderateWeek) AF VISIT 2: Moderate physical activity (last week) KCAL/day

ge6_12s_2 (intenseWeek) AF VISIT 2: Intense physical activity (last week) KCAL/day

gehos_2 (homeWorkWeek) AF VISIT 2: Physical Activity home (last week) KCAL/day

getots_2 (totalWeek) AF VISIT 2: total physical activity (last week) KCAL/day

ge2_3_5a_2 (lightYear) AF VISIT 2: light physical activity (last year) KCAL/day

ge4_5_5a_2 (moderateYear) AF VISIT 2: Moderate physical activity (last year) KCAL/day

ge6_12a_2 (intenseYear) AF VISIT 2: Intense physical activity (last year) KCAL/day

gehoa_2 (homeWorkYear) AF VISIT 2: Physical Activity home (last year) KCAL/day

getota_2 (totalYear) AF VISIT 2: total physical activity (last year) KCAL/day

tyr0 (tyrosol) Tyrosol in urine at baseline

tyr3 (tyrosol) Tyrosol in urine at 3 months

ohtyr0 (ohtyrosol) OH-tyrosol in urine at baseline

ohtyr3 (ohtyrosol) OH-tyrosol in urine at 3 months

d_ohtyr difference OH-Tyrosol

d_tyr difference Tyrosol

filter id = 400003 (FILTER)

VAR00001 (dietGroup) GENERAL: Group assigned

groupos_2 (dietGroupMed) med diet intervention yes = 1 no = 2

outliers ids = 400083, 400021, 400015, 400030, 400004, 400003 (FILTER)

VAR00004 (id) Identifier

EDAD (Age) Age

CAedad_44 (Age_cat) AGE categorized according to the median (44 years)

LDL_0 (LDL_1) Baseline LDL (mg / dl)

LDL_3 (LDL_2) LDL in 3 months (mg / dl)

d_LDL Difference LDL after-LDL before

maspirina1 (medAspirin_1) GENERAL: In the past month have you taken aspirin, Adiro or similar?

mdolor1 (medNSAID_1) GENERAL: In the past month have you taken other drugs to relieve pain or fever?

mansiedad1 (medAnxiolytic_1) GENERAL: In the past month have you taken tranquilizers, sedatives, anxiety pills or sleeping pills?

mvitMin1 (medVinMin_1) GENERAL: In the past month have you taken vitamin or mineral?

mtension1 (medHeart_1) GENERAL: In the past month have you taken medication for high blood pressure?

mcolesterol1 (medCholesterol_1) GENERAL: In the past month have you taken cholesterol medication?

mhormo1 (medHormon_1) GENERAL: In the past month have you taken hormone therapy? (Only women)

motros1 (medOther_1) GENERAL: In the past month have you taken any other medicines?

maspirina2 (medAspirin_2) FOLLOW VISIT 2: In the past month have you taken aspirin, Adiro or similar? 3 MONTHS

mdolor2 (medNSAID_2) FOLLOW VISIT 2: In the past month have you taken other drugs to relieve pain or fever? 3 MONTHS

mansiedad2 (medAnxiolytic_2) FOLLOW VISIT 2: In the past month have you taken tranquilizers, sedatives, anxiety pills or sleeping pills? 3 MONTHS

mvitMin2 (medVinMin_2) FOLLOW VISIT 2: In the past month have you taken vitamin or mineral? 3 MONTHS

mtension2 (medHeart_2) FOLLOW VISIT 2: In the past month have you taken medication for high blood pressure? 3 MONTHS

mcolesterol2 (medCholesterol_2) FOLLOW VISIT 2: In the past month have you taken cholesterol medication? 3 MONTHS

mhormo2 (medHormon_2) FOLLOW VISIT 2: In the past month have you taken hormone therapy? (Women only) 3 MONTHS

motros2 (medOther_2) FOLLOW VISIT 2: In the past month have you taken any other medicines? 3 MONTHS

menopausia_ bin (hasMenopause) GENERAL: Do you have Menopause?

Appendix B

Additional Information of Clustering Interpretations

B.1 Addition Information about the Characterization of partition P_{C_o}

B.1.1 Descriptors of NCI-IMS for P_{C_o}

	sexo (gender) - Woman	EDAD (Age)	altura1 (height)	peso1 (weight)	imc1 (BMI)	cint1 (waist)	pas_esds_2 (systolic Pressure)	pad_esds_2 (diastolic Pressure)	fc_d_2 (heart Rate)
M	↓		↑	↑	↑	↑	↑	↑	↓
YW	↑	↓	↓	↓	↓	↓	↓	↓	
WM	↑	↑	↓				↑	↑	↑

Table B.1: Biometrics characterisitcs $\sim$ Pre Baseline Classes P_{C_o} (NCI-IMS)

	tabaco (tobacco) - ex>5	tabaco (tobacco) - ex0-1	tabaco (tobacco) - ex1-5	tabaco (tobacco) - Never	tabaco (tobacco) - Yes
M					
YW					
WM	↑				↓

Table B.2: Tobacco characterisitcs ∼ Pre Baseline Classes P_{C_o} (NCI-IMS)

	est_civi (civilState) - casado/a	est_civi (civilState) - Divorciado/a	est_civi (civilState) - Separado/a	est_civi (civilState) - soltero	est_civi (civilState) - Viudo/a
M					
YW					
WM		↑		↓	↑

Table B.3: Sociodemographic characterisitcs ∼ Pre Baseline Classes P_{C_o} (NCI-IMS)

	menopausia_bin (menopause) - TRUE	pertensa (stressful)	depre (depression) - TRUE	cancer (cancer) - TRUE	fractura (bone fracture) - TRUE	disnea (dyspnea) - TRUE
M	↓	↓	↓	↓	↓	↓
YW	↓	↑	↓	↓	↓	↓
WM	↑		↓	↑	↓	↓

Table B.4: Menopause & Diseases characterisitcs $\sim$ Pre Baseline Classes P_{C_o} (NCI-IMS)

	maspirina1 (aspirin) - TRUE	mdolor1 (NSAID) - TRUE	mansiedad1 (anxiolytic) - TRUE	mvitMin1 (vitamin or minerals) - TRUE	mtension1 (heart) - TRUE	mhormo1 (hormon) - TRUE	motros1 (other) - TRUE
M	↓	↓	↓	↓	↓	↓	↓
YW	↓	↓	↓	↓	↓	↓	
WM	↓	↑	↓	↓	↓	↓	↓

Table B.5: Drugs characterisitcs $\sim$ Pre Baseline Classes P_{C_o} (NCI-IMS)

Table B.6: Biomarkers characterisitcs ~ Pre Baseline Classes P_{C_0} (NCI-IMS)

WM	YW	M	
→	←		gluc0 (glucose)
→			cholest0 (cholesterol)
→			LDL_0 (LDL)
	→	←	hdl0 (HDL)
→	←	→	tryg0 (triglycerides)
→			oxldl0 (Oxidized LDL)
←		→	isoprostanos_0 (F2 α Isoprostanes)
			ifn_g_0 (Interferon-γ)
		→	mcp_1_0 (Monocyte Chemotactic Protein-1)
←			s_p_selectin_0 (sP-selectin)
←			s_cd40l_0 (sCD40 Ligand)
			pcr0 (C-Reactive Protein)
			oxo_gg_0 (8-Oxoguanine)
			tnf_a_0 (Tumor necrosis factor-α)
	←	→	t_pa_0 (Tissue Plasminogen Activator)
			s_vcam_1_0 (Soluble cell adhesion molecules-1)
			il_6_0 (Interleukin 6)
→			il_8_0 (Interleukin 8)
→			il_10_0 (Interleukin 10)
←			tyru0 (Tyrosol)
			ohtyru0 (OHTyrosol)
			mohtyu0 (MOH_Tyrosol)

B.1.2 Automatic Profiles of P_{C_o}

M : Less Woman (-H) in *sexo* (gender); *altura1* (height) is higher(H); *peso1* (weight) is higher(H); *imc1* (BMI) is higher(H); *cint1* (waist) is higher(H); *pas_ esds_ 2* (systolic Pressure) is higher(H); *pad_ esds_ 2* (diastolic Pressure) is higher(HM); *fc_ d_ 2* (heart Rate) is lower(-H); Less TRUE (-H) in *menopausia_ bin* (menopause); *pertensa* (stressful) is lower(-H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea* (dyspnea); Less TRUE (-H) in *maspirina1* (aspirin); Less TRUE (-H) in *mdolor1* (NSAID); Less TRUE (-H) in *mansiedad1* (anxiolytic); Less TRUE (-H) in *mvitMin1* (vitamin or minerals); Less TRUE (-H) in *mtension1* (heart); Less TRUE (-H) in *mhormo1* (hormon); Less TRUE (-H) in *motros1* (other); *gluc0* (glucose) is higher(); *cholest0* (cholesterol) is lower(); *hdl0* (HDL) is lower(-H); *tryg0* (triglycerides) is higher(H); *isoprostanos_ 0* (F2 α Isoprostanes) is higher(HM); *mcp_ 1_ 0* (Monocyte Chemotactic Protein-1) is higher(H); *t_ pa_ 0* (Tissue Plasminogen Activator) is higher(H);

YW : More Woman (H) in *sexo* (gender); *EDAD* (Age) is lower(-HW); *altura1* (height) is lower(-H); *peso1* (weight) is lower(-H); *imc1* (BMI) is lower(-HM); *cint1* (waist) is lower(-H); *pas_ esds_ 2* (systolic Pressure) is lower(-H); *pad_ esds_ 2* (diastolic Pressure) is lower(-H); Less TRUE (-H) in *menopausia_ bin* (menopause); *pertensa* (stressful) is higher(H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea* (dyspnea); Less TRUE (-H) in *maspirina1* (aspirin); Less TRUE (-H) in *mdolor1* (NSAID); Less TRUE (-H) in *mansiedad1* (anxiolytic); Less TRUE (-H) in *mvitMin1* (vitamin or minerals); Less TRUE (-H) in *mtension1* (heart); Less TRUE (-H) in *mhormo1* (hormon); *gluc0* (glucose) is lower(-H); *hdl0* (HDL) is higher(HM); *tryg0* (triglycerides) is lower(-H); *oxldl0* (Oxidized LDL) is lower(); *t_ pa_ 0* (Tissue Plasminogen Activator) is lower(-H);

WM : More Woman (H) in *sexo* (gender); *EDAD* (Age) is higher(H); *altura1* (height) is lower(-H); *peso1* (weight) is lower(); *pas_ esds_ 2* (systolic Pressure) is higher(HM); *pad_ esds_ 2* (diastolic Pressure) is higher(H); *fc_ d_ 2* (heart Rate) is higher(H); More ex>5 (H) and less Yes (-HM) in *tabaco* (tobacco); More Divorciado/a (H) and less soltero (-H) and more Viudo/a (HM) in *est_ civi* (civilState); More TRUE (H) in *menopausia_ bin* (menopause); Less TRUE (-H) in *depre* (depression); More TRUE (H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea* (dyspnea); Less TRUE (-H) in *maspirina1* (aspirin); More TRUE (H) in *mdolor1* (NSAID); Less TRUE (-H) in *mansiedad1* (anxiolytic); Less TRUE (-H) in *mvitMin1* (vitamin or minerals); Less TRUE (-H) in *mtension1* (heart); Less TRUE (-H) in *mhormo1* (hormon); Less TRUE (-H) in *motros1* (other); *gluc0* (glucose) is higher(H); *cholest0* (cholesterol) is higher(H); *LDL_ 0* (LDL) is higher(H); *tryg0* (triglycerides) is higher(HM); *oxldl0* (Oxidized LDL) is higher(H); *isoprostanos_ 0* (F2 α Isoprostanes) is lower(-H); *mcp_ 1_ 0* (Monocyte Chemotactic Protein-1) is lower(); *s_ p_ selectin_ 0* (sP-selectin) is lower(-H); *s_ cd40l_ 0* (sCD40 Ligand) is lower(-H); *il_ 8_ 0* (Interleukin 8) is higher(H); *il_ 10_ 0* (Interleukin 10) is higher(H); *tyru0* (Tyrosol) is lower(-H);

B.2 Addition Information about the Characterization of partition P_{H_o}

B.2.1 Descriptors of CI-IMS for P_{H_o}

	p14_1_1 (mainOliveOil) - yes	p14_2_1 (oliveOil) - >=4spoon	p14_3_1 (vegetables) - >=2day	p14_4_1 (fruit) - >=3day	p14_5_1 (redMeat) - >=1day	p14_6_1 (butter) - >=1day	p14_7_1 (gasDrinks) - >=1day	p14_8_1 (wine) - >=7glass/week	p14_9_1 (legume) - >=3week	p14_10_1 (fish) - >=3week	p14_11_1 (commercialBakery) - >=2week	p14_12_1 (nuts) - >=3week	p14_13_1 (whiteMeat) - yes	p14_14_1 (sauce) - >=2week
WMbased	$\top$ B	$\overline{\downarrow \text{R}}$	$\uparrow$**R**	$\downarrow$**R**	$\perp$ B	$\perp$ B	$\perp$ B	$\perp$ B	$\perp$ B	$\uparrow$R	$\downarrow$R	$\overline{\downarrow \text{R}}$	$\top$ B	$\top$ B
WMwSugars	$\top$ B	$\uparrow$W	$\downarrow$R	$\uparrow$**R**	$\downarrow$**R**	$\downarrow$**M**	$\perp$ B	$\overline{\uparrow \text{W}}$	$\overline{\uparrow \text{R}}$	$\overline{\uparrow \text{R}}$	$\uparrow$**R**	$\uparrow$R	$\top$ B	$\top$ B
UH	$\top$ **R**	$\downarrow$W	$\downarrow$**R**	$\overline{\uparrow \text{R}}$	$\uparrow$R	$\uparrow$R	$\uparrow$R	$\perp$ B	$\overline{\uparrow \text{W}}$	$\downarrow$**R**	$\overline{\uparrow \text{R}}$	$\downarrow$**R**	$\top$ B	$\top$ B

Table B.7: Descriptors for partition P_{H_o} of Diet characteristics

	ge2_3_5s_1 (lightWeek)	ge4_5_5s_1 (moderateWeek)	ge6_12s_1 (intenseWeek)	gehos_1 (homeWorkWeek)	getots_1 (totalWeek)	ge2_3_5a_1 (lightYear)	ge4_5_5a_1 (moderateYear)	ge6_12a_1 (intenseYear)	gehoa_1 (homeWorkYear)	getota_1 (totalYear)
WMbased	$\overline{\uparrow \text{R}}$	$\overline{\downarrow \text{R}}$	$\overline{\uparrow \text{R}}$	$\overline{\downarrow \text{R}}$	$\overline{\uparrow \text{R}}$	$\overline{\downarrow \text{R}}$	$\overline{\downarrow \text{R}}$	$\overline{\uparrow \text{R}}$	$\overline{\downarrow \text{R}}$	$\overline{\downarrow \text{R}}$
WMwSugars	$\overline{\uparrow \text{R}}$	↑R	$\overline{\uparrow \text{R}}$	$\overline{\uparrow \text{R}}$	↑R	$\overline{\uparrow \text{R}}$	↑R	$\overline{\uparrow \text{R}}$	$\overline{\uparrow \text{R}}$	↑R
UH	↓W	↓M	↓W	$\overline{\uparrow \text{R}}$	↓R	$\overline{\downarrow \text{R}}$	↓R	↓W	$\overline{\uparrow \text{R}}$	↓R

Table B.8: Descriptors for partition P_{H_o} of Physical activity characteristics

B.2.2 Class Panel Graphs of partition P_{H_o}

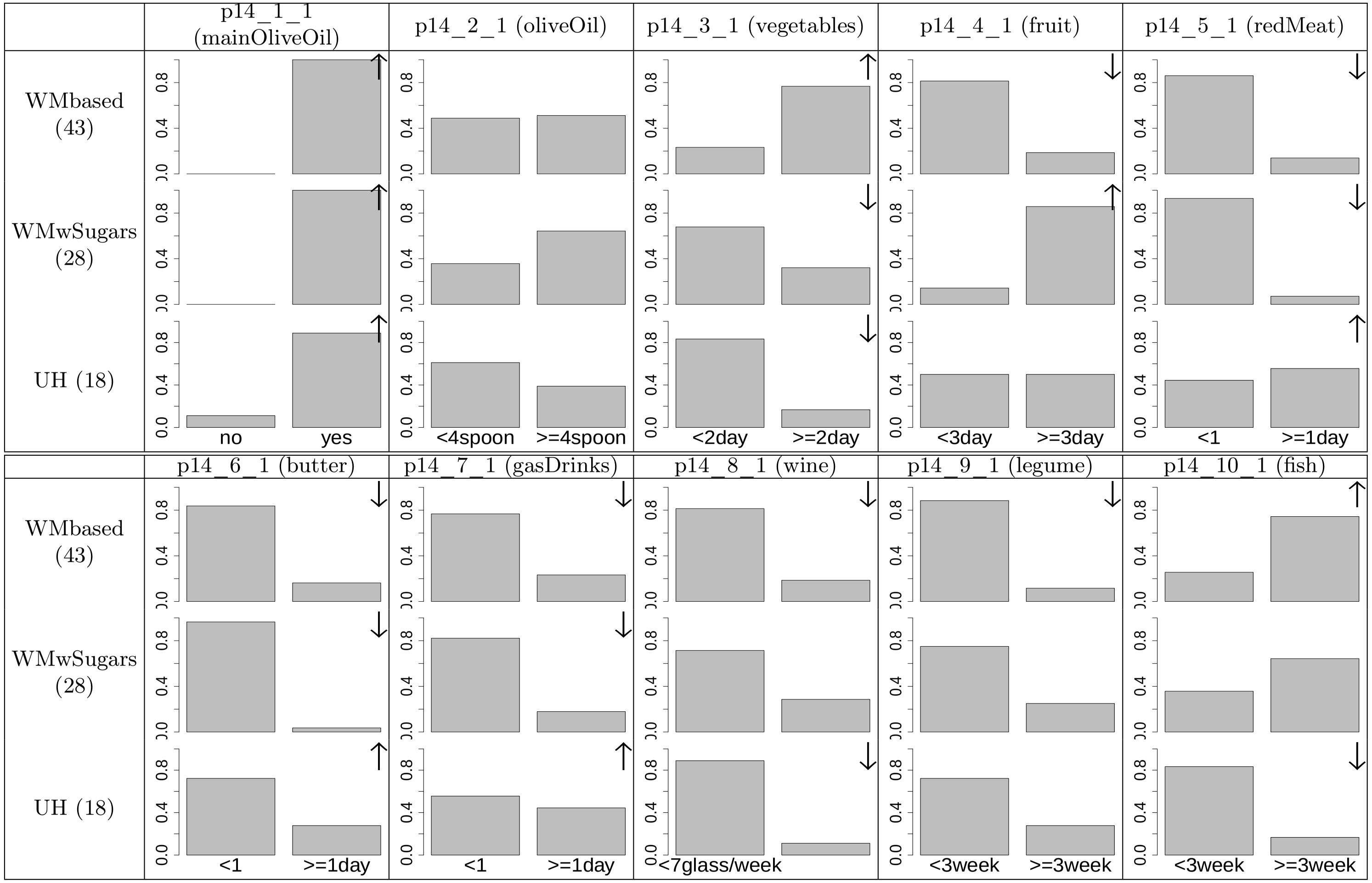


Table B.9: Diet characteristics ∼ Habits (Pre) (a)

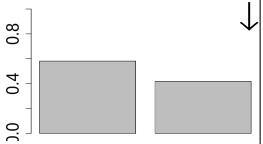
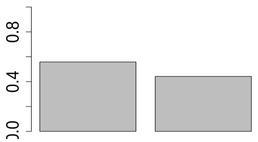
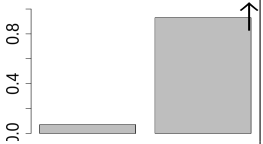
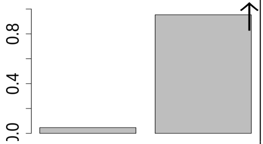
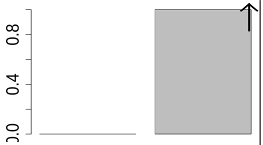
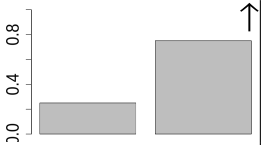
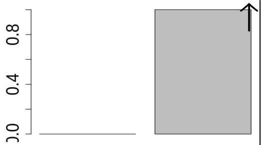
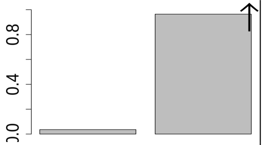
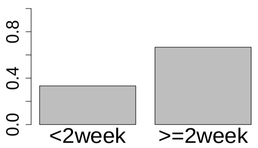
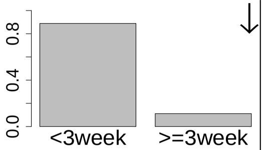
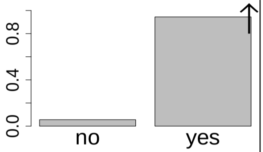
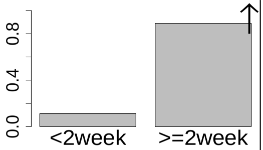
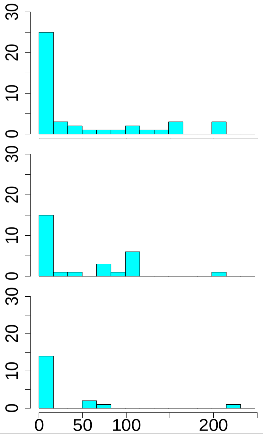
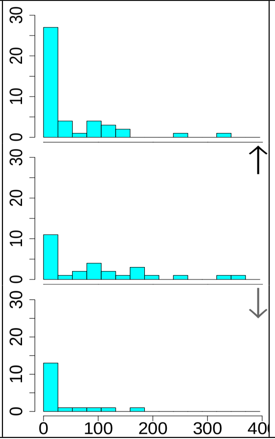
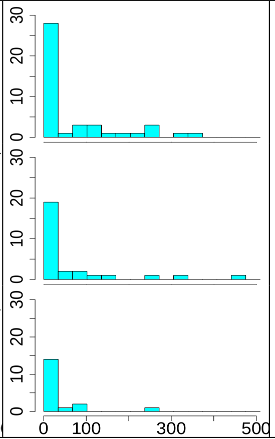
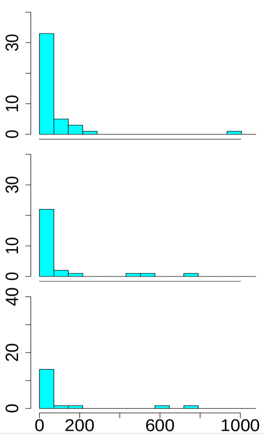
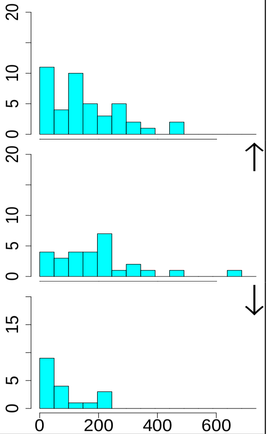
	p14_11_1 (commercialBakery)	p14_12_1 (nuts)	p14_13_1 (whiteMeat)	p14_14_1 (sauce)
WMbased (43)				
WMwSugars (28)				
UH (18)				

Table B.10: Diet characteristics ∼ Habits (Pre) (b)

	ge2_3_5s_1 (lightWeek)	ge4_5_5s_1 (moderateWeek)	ge6_12s_1 (intenseWeek)	gehos_1 (homeWorkWeek)	getots_1 (totalWeek)
WMbased (43) WMwSugars (28) UH (18)					

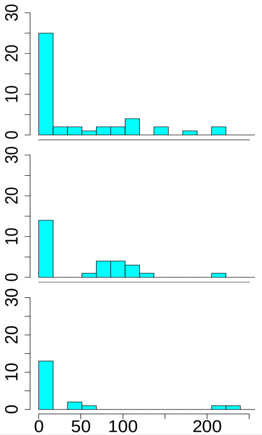
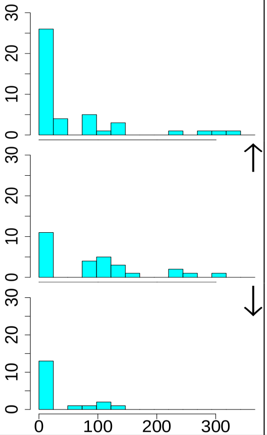
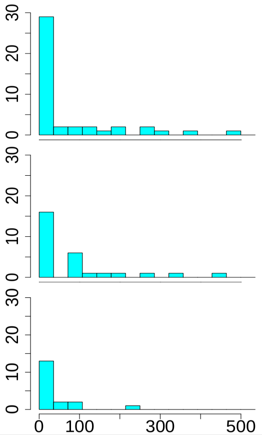
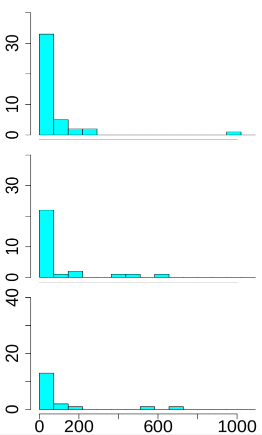
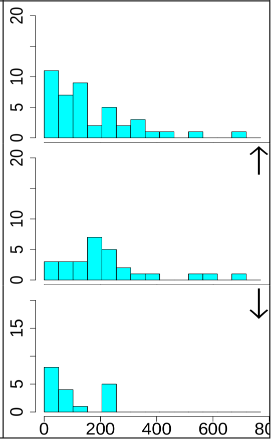
	ge2_3_5a_1 (lightYear)	ge4_5_5a_1 (moderateYear)	ge6_12a_1 (intenseYear)	gehoa_1 (homeWorkYear)	getota_1 (totalYear)
WMbased (43) WMwSugars (28) UH (18)					

Table B.11: Physical activity characteristics ~ Habits (Pre)

B.2.3 Automatic Profiles of P_{H_o}

WMbased : More yes (H) in *p14_1_1* (mainOliveOil); More >=2day (H) in *p14_3_1* (vegetables); Less >=3day (-H) in *p14_4_1* (fruit); Less >=1day (-H) in *p14_5_1* (redMeat); Less >=1day (-H) in *p14_6_1* (butter); Less >=1day (-H) in *p14_7_1* (gasDrinks); Less >=7glass/week (-H) in *p14_8_1* (wine); Less >=3week (-H) in *p14_9_1* (legume); More >=3week (H) in *p14_10_1* (fish); Less >=2week (-H) in *p14_11_1* (commercialBakery); More yes (H) in *p14_13_1* (whiteMeat); More >=2week (H) in *p14_14_1* (sauce);

WMwSugars : More yes (H) in *p14_1_1* (mainOliveOil); More >=4spoon () in *p14_2_1* (oliveOil); Less >=2day (-H) in *p14_3_1* (vegetables); More >=3day (H) in *p14_4_1* (fruit); Less >=1day (-H) in *p14_5_1* (redMeat); Less >=1day (-H) in *p14_6_1* (butter); Less >=1day (-H) in *p14_7_1* (gasDrinks); More >=2week (H) in *p14_11_1* (commercialBakery); More >=3week (H) in *p14_12_1* (nuts); More yes (H) in *p14_13_1* (whiteMeat); More >=2week (H) in *p14_14_1* (sauce); *ge4_5_5s_1* (moderateWeek) is higher(H); *getots_1* (totalWeek) is higher(H); *ge4_5_5a_1* (moderateYear) is higher(H); *getota_1* (totalYear) is higher(H);

UH : More yes (H) in *p14_1_1* (mainOliveOil); Less >=4spoon () in *p14_2_1* (oliveOil); Less >=2day (-H) in *p14_3_1* (vegetables); More >=1day (H) in *p14_5_1* (redMeat); More >=1day (H) in *p14_6_1* (butter); More >=1day (H) in *p14_7_1* (gasDrinks); Less >=7glass/week (-H) in *p14_8_1* (wine); Less >=3week (-H) in *p14_10_1* (fish); Less >=3week (-H) in *p14_12_1* (nuts); More yes (H) in *p14_13_1* (whiteMeat); More >=2week (H) in *p14_14_1* (sauce); *ge2_3_5s_1* (lightWeek) is lower(); *ge4_5_5s_1* (moderateWeek) is lower(-HM); *ge6_12s_1* (intenseWeek) is lower(); *getots_1* (totalWeek) is lower(-H); *ge4_5_5a_1* (moderateYear) is lower(-H); *ge6_12a_1* (intenseYear) is lower(); *getota_1* (totalYear) is lower(-H);

B.3 Addition Information about the Characterization of partition P_o

B.3.1 Descriptors of CI-IMS for P_o

	sexo (gender) - Woman	EDAD (Age)	altura1 (height)	peso1 (weight)	imc1 (BMI)	cint1 (waist)	pas_esds_2 (systolic Pressure)	pad_esds_2 (diastolic Pressure)	fc_d_2 (heart Rate)
M-WMbased	↓**R**	$\overline{\uparrow \mathrm{M}}$	↑R	↑R	$\overline{\uparrow \mathrm{M}}$	↑R	↑R	↑R	↓M
M-WMwSugars	↓**R**	$\overline{\downarrow \mathrm{R}}$	↑R	↑R	↑M	↑R	↑R	↑M	$\overline{\uparrow \mathrm{R}}$
M-UH	↓**R**	↓R	↑R	↑R	↑R	↑R	↑W	$\overline{\uparrow \mathrm{R}}$	↓R
YW-WMbased	↑**R**	↓W	$\overline{\downarrow \mathrm{R}}$	↓R	↓R	↓R	↓R	↓W	$\overline{\uparrow \mathrm{R}}$
YW-WMwSugars	↑**R**	$\overline{\downarrow \mathrm{R}}$	↓R	↓R	↓M	↓R	↓M	$\overline{\downarrow \mathrm{W}}$	$\overline{\downarrow \mathrm{R}}$
YW-UH	↑**R**	↓M	↓R	$\overline{\downarrow \mathrm{M}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↓W	↓R	$\overline{\uparrow \mathrm{R}}$
WM-WMbased	↑**R**	↑R	↓R	↓R	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	↑R	↑R
WM-WMwSugars	↑**R**	↑R	↓R	$\overline{\downarrow \mathrm{R}}$	↑R	↑W	↑R	↑R	↑R

Table B.12: Descriptors for partition P_o (CI-IMS) of Biometrics characteristics

	tabaco (tobacco) - ex>5	tabaco (tobacco) - ex0-1	tabaco (tobacco) - ex1-5	tabaco (tobacco) - Never	tabaco (tobacco) - Yes
M-WMbased	⊥ **W**	⊥ B	⊥ B	$\overline{\uparrow \mathrm{R}}$	⊥ B
M-WMwSugars	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{W}}$	$\overline{\downarrow \mathrm{R}}$	↑**W**	$\overline{\downarrow \mathrm{R}}$
M-UH	⊥ B	⊥ **R**	⊥ **R**	↓R	⊥ B
YW-WMbased	⊥ B	↓**W**	⊥ B	$\overline{\downarrow \mathrm{R}}$	↑W
YW-WMwSugars	⊥ B	⊥ B	⊥ **W**	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$
YW-UH	⊥ B	⊥ **R**	⊥ B	$\overline{\uparrow \mathrm{R}}$	⊥ B
WM-WMbased	⊥ **R**	⊥ B	⊥ B	$\overline{\uparrow \mathrm{R}}$	⊥ B
WM-WMwSugars	⊥ **R**	⊥ **R**	⊥ B	$\overline{\uparrow \mathrm{R}}$	↓**R**

Table B.13: Descriptors for partition P_o (CI-IMS) of Tobacco characteristics

	est_civi (civilState) - casado/a	est_civi (civilState) - Divorciado/a	est_civi (civilState) - Separado/a	est_civi (civilState) - soltero	est_civi (civilState) - Viudo/a
M-WMbased	$\top$ B	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{M}}$	$\downarrow$W	$\uparrow$R
M-WMwSugars	$\top$ B	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$
M-UH	$\downarrow$W	$\perp$ B	$\perp$ B	$\uparrow$R	$\perp$ B
YW-WMbased	$\overline{\downarrow \mathrm{R}}$	$\perp$ B	$\perp$ B	$\overline{\uparrow \mathrm{R}}$	$\perp$ B
YW-WMwSugars	$\overline{\downarrow \mathrm{R}}$	$\perp$ B	$\perp$ B	$\overline{\uparrow \mathrm{M}}$	$\perp$ B
YW-UH	$\overline{\downarrow \mathrm{R}}$	$\perp$ B	$\perp$ **M**	$\perp$ B	$\perp$ B
WM-WMbased	$\boldsymbol{\uparrow}$**M**	$\uparrow$R	$\overline{\downarrow \mathrm{M}}$	$\downarrow$R	$\overline{\downarrow \mathrm{R}}$
WM-WMwSugars	$\overline{\uparrow \mathrm{R}}$	$\perp$ B	$\perp$ **R**	$\boldsymbol{\downarrow}$**R**	$\perp$ **R**

Table B.14: Descriptors for partition P_o (CI-IMS) of Sociodemographic characteristics

	menopausia_bin (menopause) - TRUE	pertensa (stressful)	depre (depression) - TRUE	cancer (cancer) - TRUE	fractura (bone fracture) - TRUE	disnea (dyspnea) - TRUE
M-WMbased	**↓R**	↓R	⊥ B	⊥ B	⊥ B	⊥ B
M-WMwSugars	**↓R**	↓R	⊥ B	⊥ B	↑R	↑R
M-UH	**↓R**	↓R	⊥ B	⊥ B	⊥ B	⊥ B
YW-WMbased	**↓R**	↑W	⊥ B	⊥ B	⊥ B	⊥ B
YW-WMwSugars	**↓M**	$\overline{\uparrow \text{R}}$	⊥ **R**	⊥ B	⊥ B	⊥ B
YW-UH	**↓R**	↑R	⊥ B	⊥ B	⊥ B	⊥ B
WM-WMbased	**↑R**	↑M	⊥ B	↑R	⊥ B	⊥ B
WM-WMwSugars	**↑R**	$\overline{\downarrow \text{R}}$	⊥ B	⊥ **R**	⊥ B	⊥ B

Table B.15: Descriptors for partition P_o (CI-IMS) of Menopause & Diseases characteristics

	maspirina1 (aspirin) - TRUE	mdolor1 (NSAID) - TRUE	mansiedad1 (anxiolytic) - TRUE	mvitMin1 (vitamin or minerals) - TRUE	mtension1 (heart) - TRUE	mhormo1 (hormon) - TRUE	motros1 (other) - TRUE
M-WMbased	⊥ **R**	↓**R**	↓**M**	↓**W**	⊥ B	⊥ B	↓**W**
M-WMwSugars	⊥ B	↓**R**	↓**W**	↓**W**	⊥ **R**	⊥ B	↑R
M-UH	⊥ **M**	↓**R**	↓**M**	↓**W**	⊥ B	⊥ B	↓**W**
YW-WMbased	⊥ B	↑R	⊥ B	⊥ B	↓**W**	⊥ B	$\overline{\uparrow \text{M}}$
YW-WMwSugars	⊥ B	⊥ B	⊥ **M**	⊥ B	⊥ B	⊥ B	⊥ B
YW-UH	⊥ B	↓**R**	⊥ B	↓**W**	⊥ B	⊥ **R**	↑W
WM-WMbased	⊥ B	$\overline{\uparrow \text{W}}$	↓**M**	⊥ **W**	⊥ **R**	⊥ **R**	↑R
WM-WMwSugars	⊥ **R**	↑R	↓**M**	↓**W**	⊥ **R**	⊥ B	↓**R**

Table B.16: Descriptors for partition P_o (CI-IMS) of Drugs characteristics

	gluc0 (glucose)	cholest0 (cholesterol)	LDL_0 (LDL)	hdl0 (HDL)	tryg0 (triglycerides)	oxldl0 (Oxidized LDL)	isoprostanos_0 (F2 α Isoprostanes)	ifn_g_0 (Interferon-γ)	mcp_1_0 (Monocyte Chemotactic Protein-1)	s_p_selectin_0 (sP-selectin)	s_cd40l_0 (sCD40 Ligand)	pcr0 (C-Reactive Protein)	oxo_gg_0 (8-Oxoguanine)	tnf_a_0 (Tumor necrosis factor-α)	t_pa_0 (Tissue Plasminogen Activator)	s_vcam_1_0 (Soluble cell adhesion molecules-1)	il_6_0 (Interleukin 6)	il_8_0 (Interleukin 8)	il_10_0 (Interleukin 10)	tyru0 (Tyrosol)	ohtyru0 (OHTyrosol)	mohtyu0 (MOH_Tyrosol)
M-WMbased	$\overline{\uparrow W}$	$\overline{\downarrow R}$	$\overline{\uparrow R}$	↓R	$\overline{\uparrow R}$	$\overline{\uparrow R}$	↑R	$\overline{\downarrow R}$	↑R	$\overline{\downarrow R}$	$\overline{\downarrow W}$	$\overline{\downarrow R}$	$\overline{\uparrow R}$	$\overline{\downarrow R}$	$\overline{\uparrow R}$	$\overline{\downarrow M}$	↓M	$\overline{\downarrow R}$	↓M	↑R	↑R	↑R
M-WMwSugars	↑R	$\overline{\downarrow W}$	↓M	↓R	↑R	$\overline{\downarrow R}$	$\overline{\uparrow R}$	$\overline{\downarrow R}$	↑R	↑R	↓R	↓W	$\overline{\uparrow M}$	$\overline{\downarrow R}$	↑W	$\overline{\downarrow M}$	$\overline{\downarrow R}$	$\overline{\downarrow R}$	$\overline{\uparrow R}$	↓M	$\overline{\downarrow R}$	$\overline{\downarrow W}$
M-UH	$\overline{\uparrow R}$	↓R	↓R	↓R	↑R	$\overline{\downarrow R}$	↑W	$\overline{\downarrow R}$	$\overline{\uparrow W}$	↑M	↑R	$\overline{\downarrow R}$	↓M	$\overline{\downarrow R}$	↑R	$\overline{\downarrow M}$	$\overline{\downarrow R}$	$\overline{\downarrow R}$	$\overline{\downarrow R}$	$\overline{\downarrow W}$	$\overline{\downarrow M}$	$\overline{\downarrow R}$
YW-WMbased	↓M	$\overline{\uparrow R}$	$\overline{\uparrow R}$	↑R	↓W	$\overline{\downarrow R}$	$\overline{\downarrow R}$	$\overline{\downarrow R}$	↓M	$\overline{\uparrow R}$	$\overline{\uparrow R}$	↑M	$\overline{\downarrow R}$	$\overline{\downarrow R}$	$\overline{\downarrow R}$	$\overline{\uparrow R}$	$\overline{\uparrow R}$	$\overline{\uparrow R}$	$\overline{\uparrow R}$	$\overline{\downarrow R}$	$\overline{\downarrow W}$	$\overline{\downarrow W}$
YW-WMwSugars	$\overline{\downarrow M}$	↓W	↓W	$\overline{\uparrow R}$	↓R	↓W	$\overline{\downarrow R}$	↑R	$\overline{\uparrow R}$	$\overline{\downarrow R}$	$\overline{\uparrow R}$	↓R	$\overline{\downarrow R}$	↑M	↓R	$\overline{\uparrow R}$	$\overline{\downarrow W}$	$\overline{\downarrow R}$	$\overline{\downarrow R}$	$\overline{\uparrow R}$	↑W	↑M
YW-UH	↓R	↓R	↓R	$\overline{\uparrow R}$	↓R	↓R	↑W	$\overline{\uparrow R}$	$\overline{\uparrow R}$	$\overline{\downarrow R}$	$\overline{\uparrow R}$	$\overline{\downarrow R}$	↑W	↑M	↓R	↑W	↑R	$\overline{\downarrow R}$	↓M	↑R	$\overline{\downarrow R}$	$\overline{\uparrow R}$
WM-WMbased	↑R	↑R	↑R	↑R	$\overline{\uparrow R}$	↑R	↓R	$\overline{\downarrow R}$	$\overline{\downarrow R}$	↓R	↓M	↑R	$\overline{\downarrow R}$	$\overline{\downarrow R}$	$\overline{\uparrow R}$	$\overline{\downarrow M}$	$\overline{\downarrow R}$	↑R	↑R	↓M	$\overline{\downarrow W}$	$\overline{\downarrow R}$
WM-WMwSugars	↑M	↑R	↑R	↓M	↑R	↑R	$\overline{\downarrow R}$	$\overline{\downarrow R}$	↓R	$\overline{\downarrow W}$	↓R	↓W	$\overline{\uparrow R}$	$\overline{\downarrow R}$	$\overline{\uparrow R}$	$\overline{\downarrow M}$	$\overline{\uparrow R}$	$\overline{\downarrow R}$	↑R	↓R	$\overline{\downarrow R}$	$\overline{\downarrow W}$

Table B.17: Descriptors for partition P_o (CI-IMS) of Biomarkers characteristics

	p14_1_1 (mainOliveOil) - yes	p14_2_1 (oliveOil) - >=4spoon	p14_3_1 (vegetables) - >=2day	p14_4_1 (fruit) - >=3day	p14_5_1 (redMeat) - >=1day	p14_6_1 (butter) - >=1day	p14_7_1 (gasDrinks) - >=1day	p14_8_1 (wine) - >=7glass/week	p14_9_1 (legume) - >=3week	p14_10_1 (fish) - >=3week	p14_11_1 (commercialBakery) - >=2week	p14_12_1 (nuts) - >=3week	p14_13_1 (whiteMeat) - yes	p14_14_1 (sauce) - >=2week
M-WMbased	⊤ B	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↓**R**	⊥ B	↑R	⊥ B	$\overline{\uparrow \mathrm{W}}$	⊥ B	$\overline{\uparrow \mathrm{R}}$	↓**R**	$\overline{\uparrow \mathrm{W}}$	⊤ **R**	⊤ B
M-WMwSugars	⊤ B	$\overline{\uparrow \mathrm{R}}$	↓**R**	↑**R**	⊥ B	↓**R**	⊥ B	↑R	↑R	$\overline{\uparrow \mathrm{R}}$	↑**R**	$\overline{\uparrow \mathrm{W}}$	⊤ B	⊤ B
M-UH	⊤ **R**	$\overline{\uparrow \mathrm{R}}$	↓**R**	$\overline{\downarrow \mathrm{R}}$	↑**R**	⊥ B	↑R	⊥ B	↑M	↓**R**	$\overline{\downarrow \mathrm{R}}$	↓**R**	⊤ **M**	⊤ B
YW-WMbased	⊤ B	$\overline{\downarrow \mathrm{R}}$	↑**R**	↓**R**	⊥ B	⊥ B	$\overline{\uparrow \mathrm{R}}$	⊥ B	⊥ B	↑**R**	↓W	$\overline{\downarrow \mathrm{R}}$	⊤ B	⊤ B
YW-WMwSugars	⊤ B	↑M	$\overline{\downarrow \mathrm{R}}$	↑**R**	↓**M**	⊥ B	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	⊥ B	$\overline{\uparrow \mathrm{R}}$	↑**R**	↑R	⊤ B	⊤ B
YW-UH	⊤ B	↓**R**	↓**R**	$\overline{\uparrow \mathrm{R}}$	↑M	↑R	⊥ B	⊥ B	⊥ B	↓**R**	↑**W**	↓**R**	⊤ B	⊤ B
WM-WMbased	⊤ B	$\overline{\uparrow \mathrm{R}}$	↑**R**	↓**R**	↓**R**	↓**R**	↓**W**	⊥ B	↓**R**	$\overline{\uparrow \mathrm{W}}$	↓R	↓R	⊤ B	⊤ B
WM-WMwSugars	⊤ B	$\overline{\uparrow \mathrm{R}}$	↓**R**	↑**R**	↓**R**	↓**R**	↓**R**	⊥ B	$\overline{\uparrow \mathrm{W}}$	$\overline{\uparrow \mathrm{W}}$	↑**R**	↑**R**	⊤ B	⊤ **R**

Table B.18: Descriptors for partition P_o (CI-IMS) of Diet characteristics

	ge2_3_5s_1 (lightWeek)	ge4_5_5s_1 (moderateWeek)	ge6_12s_1 (intenseWeek)	gehos_1 (homeWorkWeek)	getots_1 (totalWeek)	ge2_3_5a_1 (lightYear)	ge4_5_5a_1 (moderateYear)	ge6_12a_1 (intenseYear)	gehoa_1 (homeWorkYear)	getota_1 (totalYear)
M-WMbased	$\overline{\uparrow \text{R}}$	$\overline{\downarrow \text{R}}$	↑R	↓W	↑R	$\overline{\uparrow \text{R}}$	$\overline{\uparrow \text{R}}$	↑R	↓W	↑R
M-WMwSugars	↑R	↑R	↓R	↓R	↑R	↑R	↑R	↓M	↓R	↑R
M-UH	$\overline{\downarrow \text{W}}$	↓R	$\overline{\downarrow \text{R}}$	↓R	↓R	↓W	↓R	$\overline{\downarrow \text{W}}$	↓R	↓R
YW-WMbased	$\overline{\downarrow \text{R}}$	$\overline{\downarrow \text{R}}$	$\overline{\uparrow \text{R}}$	$\overline{\downarrow \text{R}}$	$\overline{\downarrow \text{R}}$	$\overline{\downarrow \text{R}}$	$\overline{\downarrow \text{R}}$	$\overline{\downarrow \text{R}}$	$\overline{\uparrow \text{R}}$	$\overline{\downarrow \text{M}}$
YW-WMwSugars	$\overline{\downarrow \text{R}}$	$\overline{\uparrow \text{W}}$	↑R	$\overline{\downarrow \text{R}}$	↑R	$\overline{\downarrow \text{R}}$	$\overline{\uparrow \text{W}}$	↑R	$\overline{\downarrow \text{R}}$	↑R
YW-UH	↓W	↓W	↓M	↑R	↓R	$\overline{\downarrow \text{R}}$	↓W	↓M	↑R	↓R
WM-WMbased	↑R	↓W	↓R	$\overline{\downarrow \text{R}}$	$\overline{\downarrow \text{W}}$	↑R	↓M	↓R	$\overline{\uparrow \text{R}}$	↓M
WM-WMwSugars	$\overline{\uparrow \text{R}}$	$\overline{\uparrow \text{R}}$	↓R	↑R	$\overline{\downarrow \text{W}}$	$\overline{\uparrow \text{M}}$	$\overline{\uparrow \text{W}}$	↓W	↑R	$\overline{\uparrow \text{R}}$

Table B.19: Descriptors for partition P_o (CI-IMS) of Physical activity characteristics

B.3.2 Class Panel Graphs of partition P_o

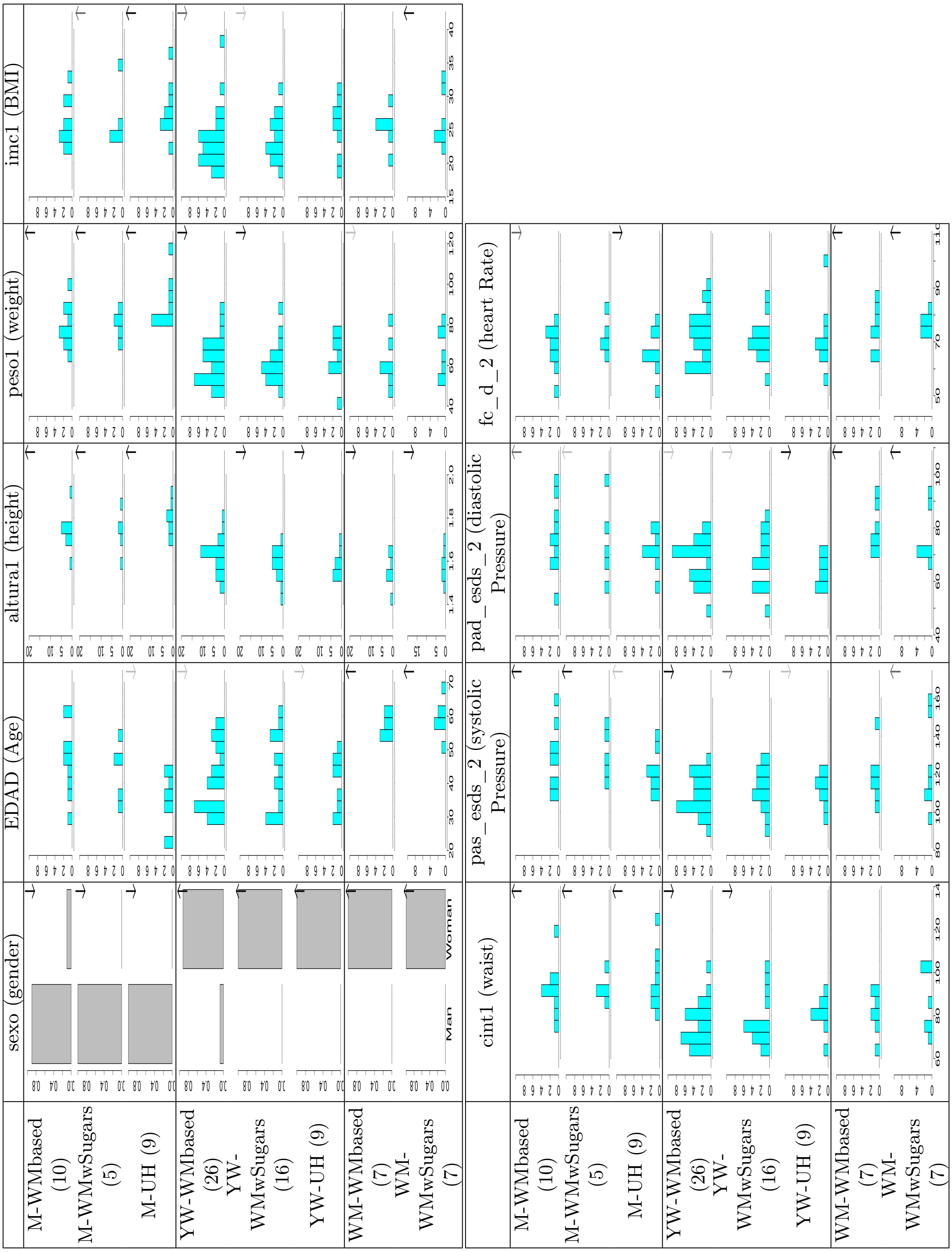


Table B.20: Biometrics characteristics ~ Cross (Pre)

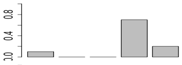
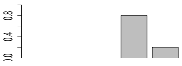
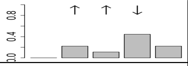
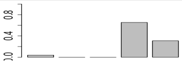
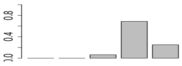
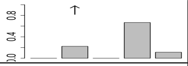
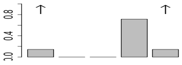
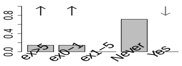
	tabaco (tobacco)
M-WMbased (10)	
M-WMwSugars (5)	
M-UH (9)	
YW-WMbased (26)	
YW-WMwSugars (16)	
YW-UH (9)	
WM-WMbased (7)	
WM-WMwSugars (7)	

Table B.21: Tobacco characteristics ∼ Cross (Pre)

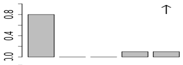
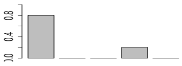
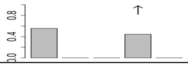
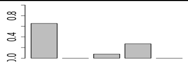
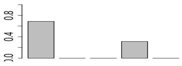
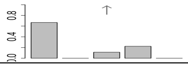
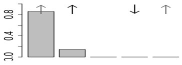
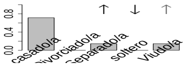
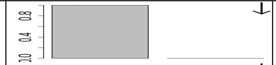
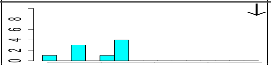
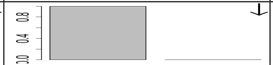
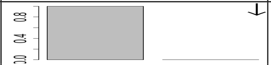
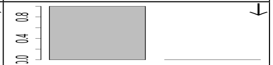
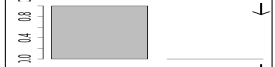
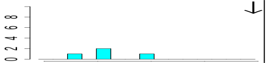
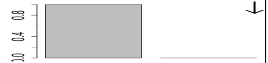
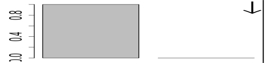
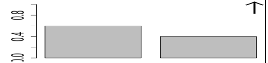
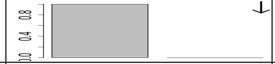
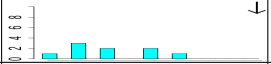
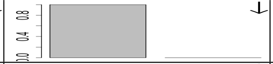
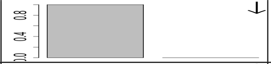
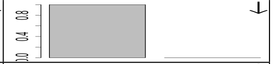
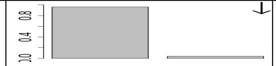
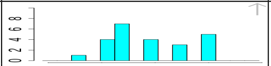
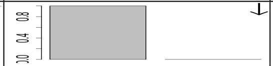
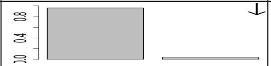
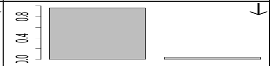
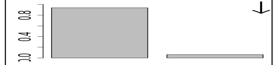
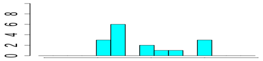
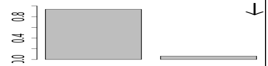
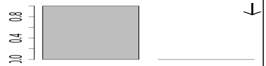
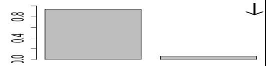
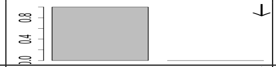
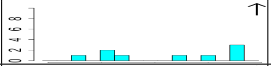
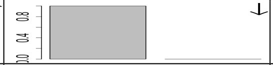
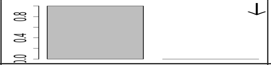
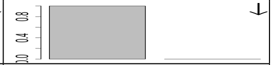
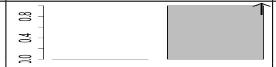
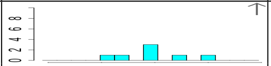
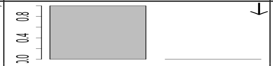
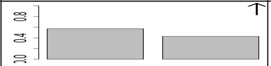
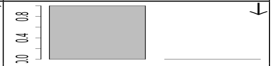
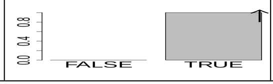
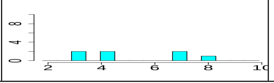
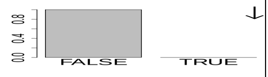
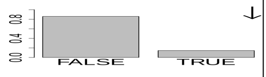
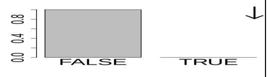
	est_civi (civilState)
M-WMbased (10)	
M-WMwSugars (5)	
M-UH (9)	
YW-WMbased (26)	
YW-WMwSugars (16)	
YW-UH (9)	
WM-WMbased (7)	
WM-WMwSugars (7)	

Table B.22: Sociodemographic characteristics ∼ Cross (Pre)

	menopausia_bin (menopause)	pertensa (stressful)	depre (depression)	cancer (cancer)	fractura (bone fracture)
M-WMbased (10)					
M-WMwSugars (5)					
M-UH (9)					
YW-WMbased (26)					
YW-WMwSugars (16)					
YW-UH (9)					
WM-WMbased (7)					
WM-WMwSugars (7)					

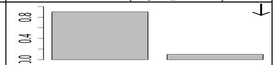
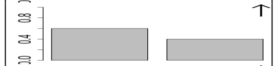
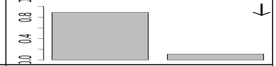
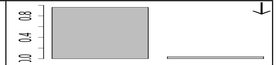
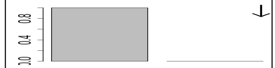
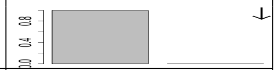
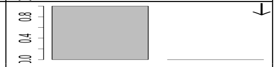
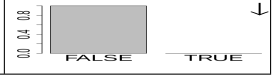
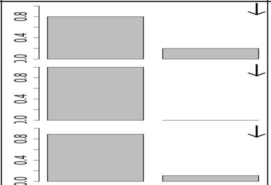
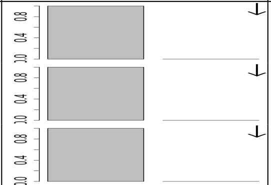
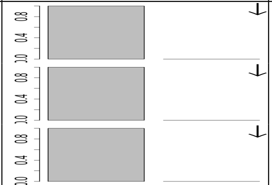
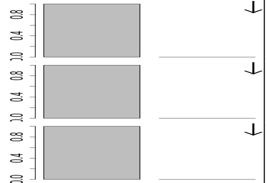
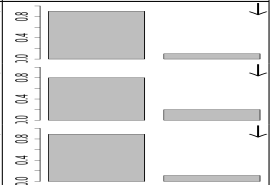
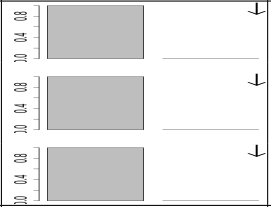
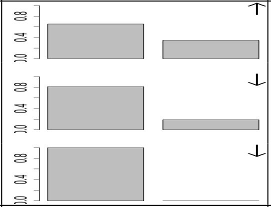
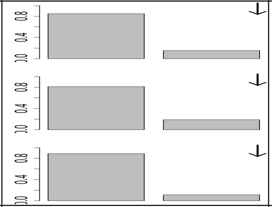
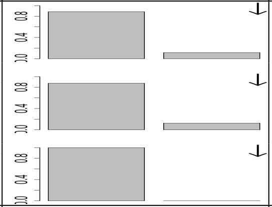
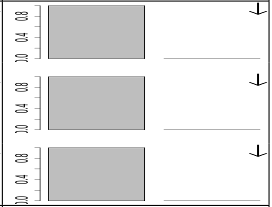
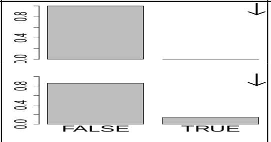
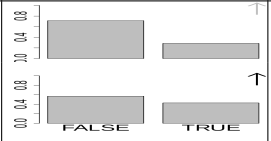
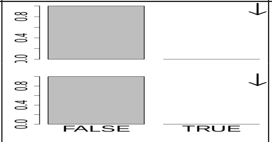
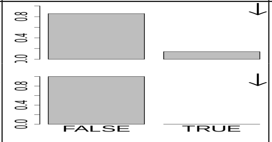
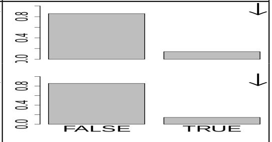
	disnea (dyspnea)
M-WMbased (10)	
M-WMwSugars (5)	
M-UH (9)	
YW-WMbased (26)	
YW-WMwSugars (16)	
YW-UH (9)	
WM-WMbased (7)	
WM-WMwSugars (7)	

Table B.23: Menopause & Diseases characteristics ∼ Cross (Pre)

	maspirina1 (aspirin)	mdolor1 (NSAID)	mansiedad1 (anxiolytic)	mvitMin1 (vitamin or minerals)	mtension1 (heart)
M-WMbased (10) M-WMwSugars (5) M-UH (9)					
YW-WMbased (26) YW-WMwSugars (16) YW-UH (9)					
WM-WMbased (7) WM-WMwSugars (7)					

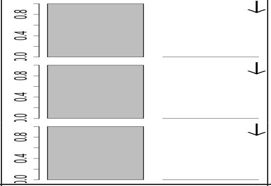
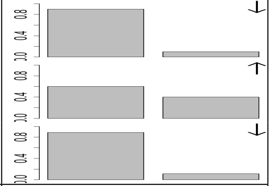
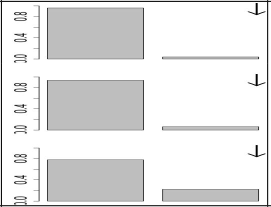
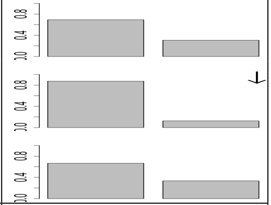
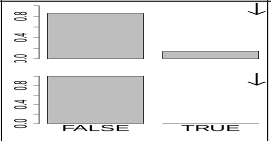
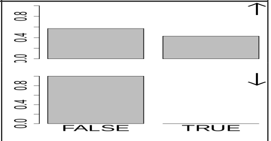
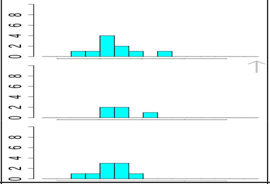
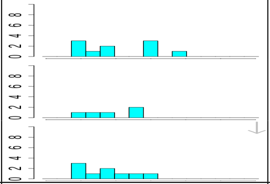
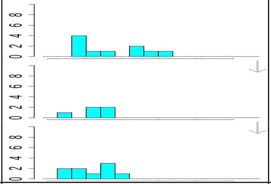
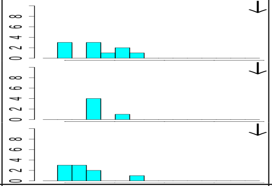
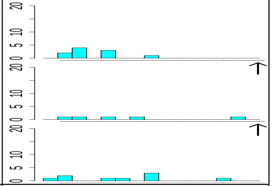
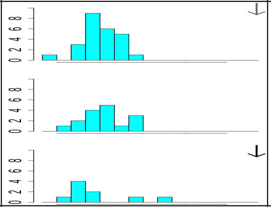
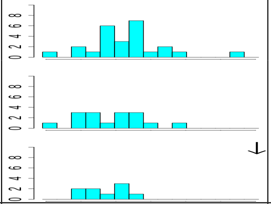
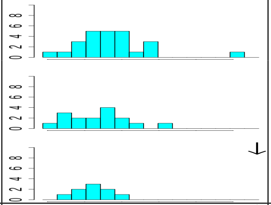
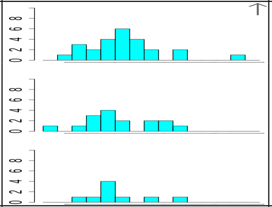
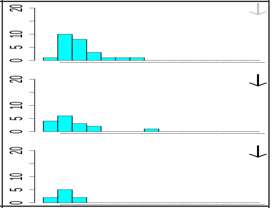
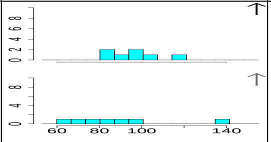
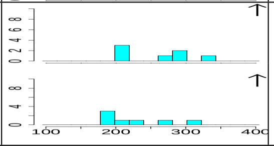
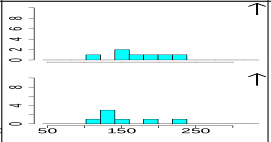
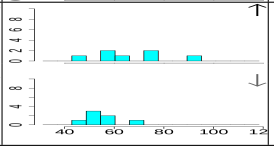
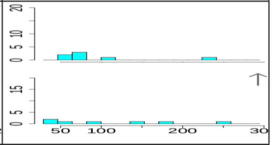
	mhormo1 (hormon)	motros1 (other)
M-WMbased (10) M-WMwSugars (5) M-UH (9)		
YW-WMbased (26) YW-WMwSugars (16) YW-UH (9)		
WM-WMbased (7) WM-WMwSugars (7)		

Table B.24: Drugs characteristics ∼ Cross (Pre)

	gluc0 (glucose)	cholest0 (cholesterol)	LDL_0 (LDL)	hdl0 (HDL)	tryg0 (triglycerides)
M-WMbased (10) M-WMwSugars (5) M-UH (9)					
YW-WMbased (26) YW-WMwSugars (16) YW-UH (9)					
WM-WMbased (7) WM-WMwSugars (7)					

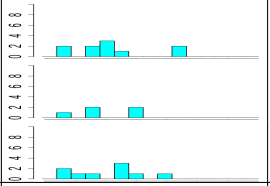
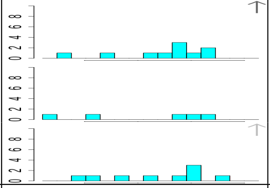
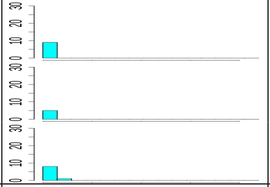
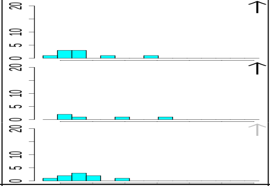
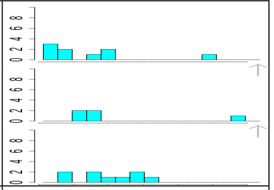
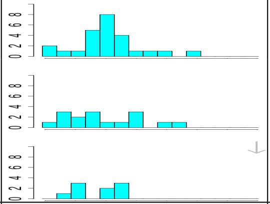
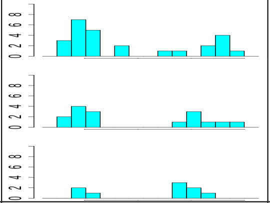
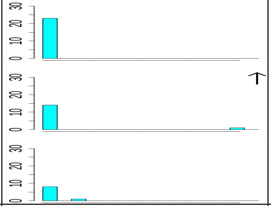
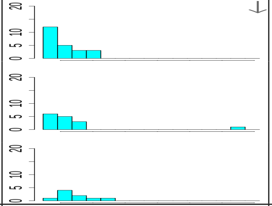
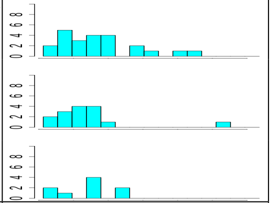
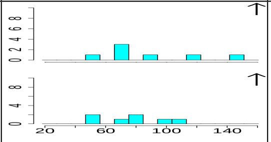
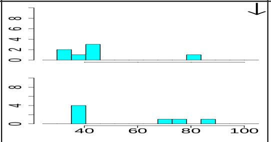
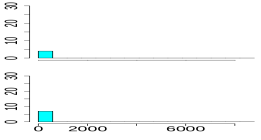
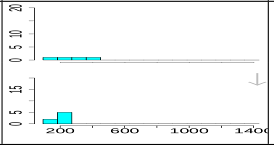
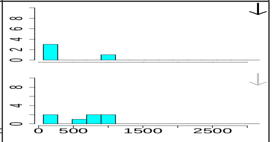
	oxldl0 (Oxidized LDL)	isoprostanos_0 (F2 α Isoprostanes)	ifn_g_0 (Interferon-γ)	mcp_1_0 (Monocyte Chemotactic Protein-1)	s_p_selectin_0 (sP-selectin)
M-WMbased (10) M-WMwSugars (5) M-UH (9)					
YW-WMbased (26) YW-WMwSugars (16) YW-UH (9)					
WM-WMbased (7) WM-WMwSugars (7)					

Table B.25: Biomarkers characteristics ~ Cross (Pre) (a)

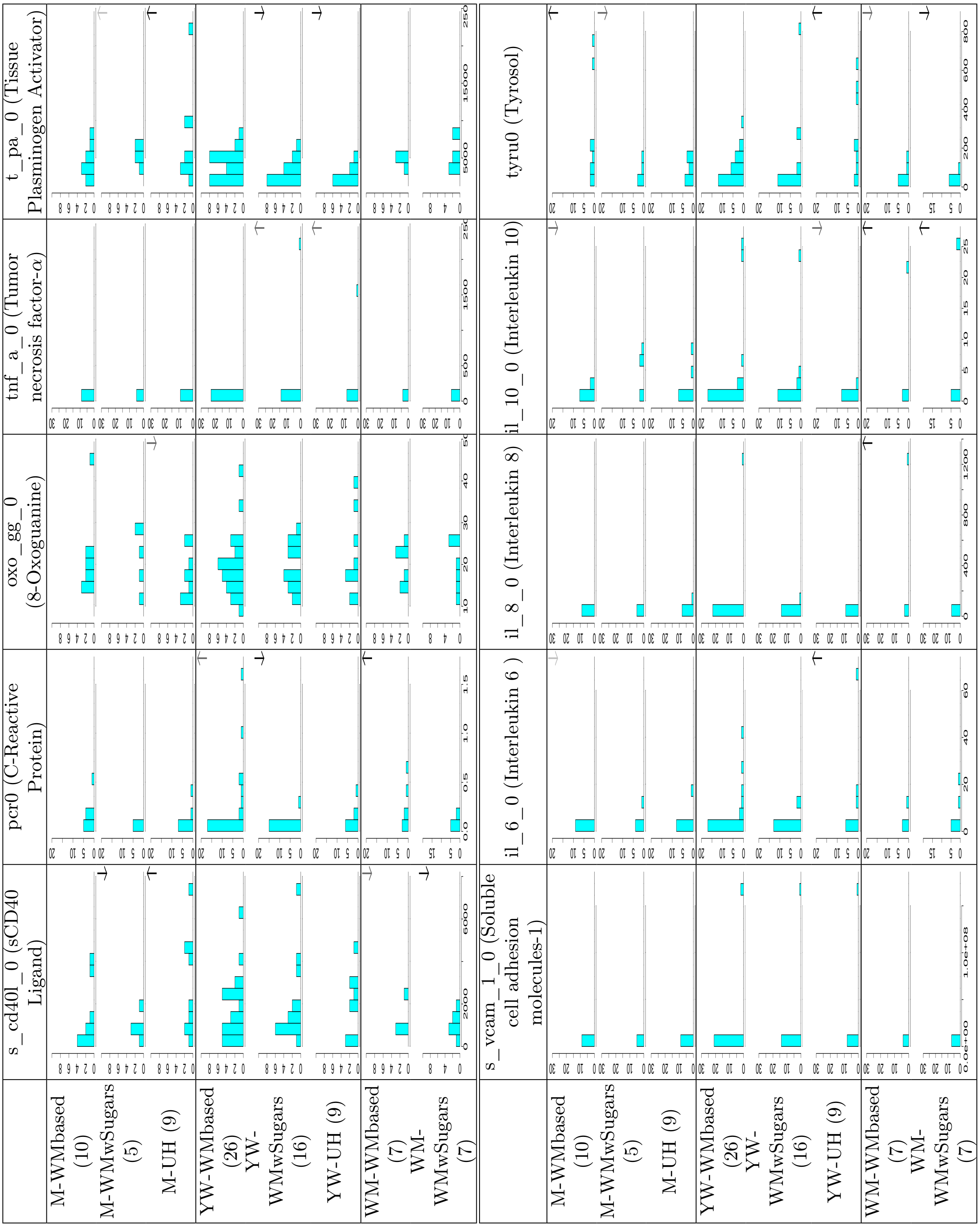


Table B.26: Biomarkers characteristics ∼ Cross (Pre) (b)

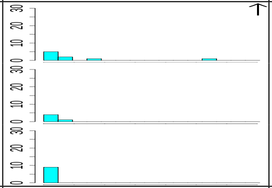
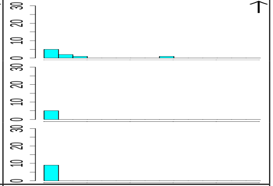
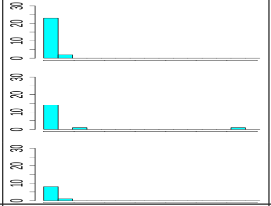
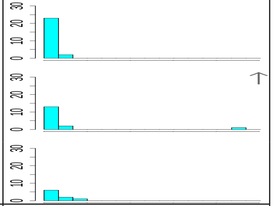
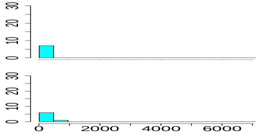
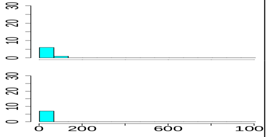
	ohtyru0 (OHTyrosol)	mohtyu0 (MOH_Tyrosol)
M-WMbased (10) M-WMwSugars (5) M-UH (9)		
YW-WMbased (26) YW-WMwSugars (16) YW-UH (9)		
WM-WMbased (7) WM-WMwSugars (7)		

Table B.27: Biomarkers characteristics ∼ Cross (Pre) (c)

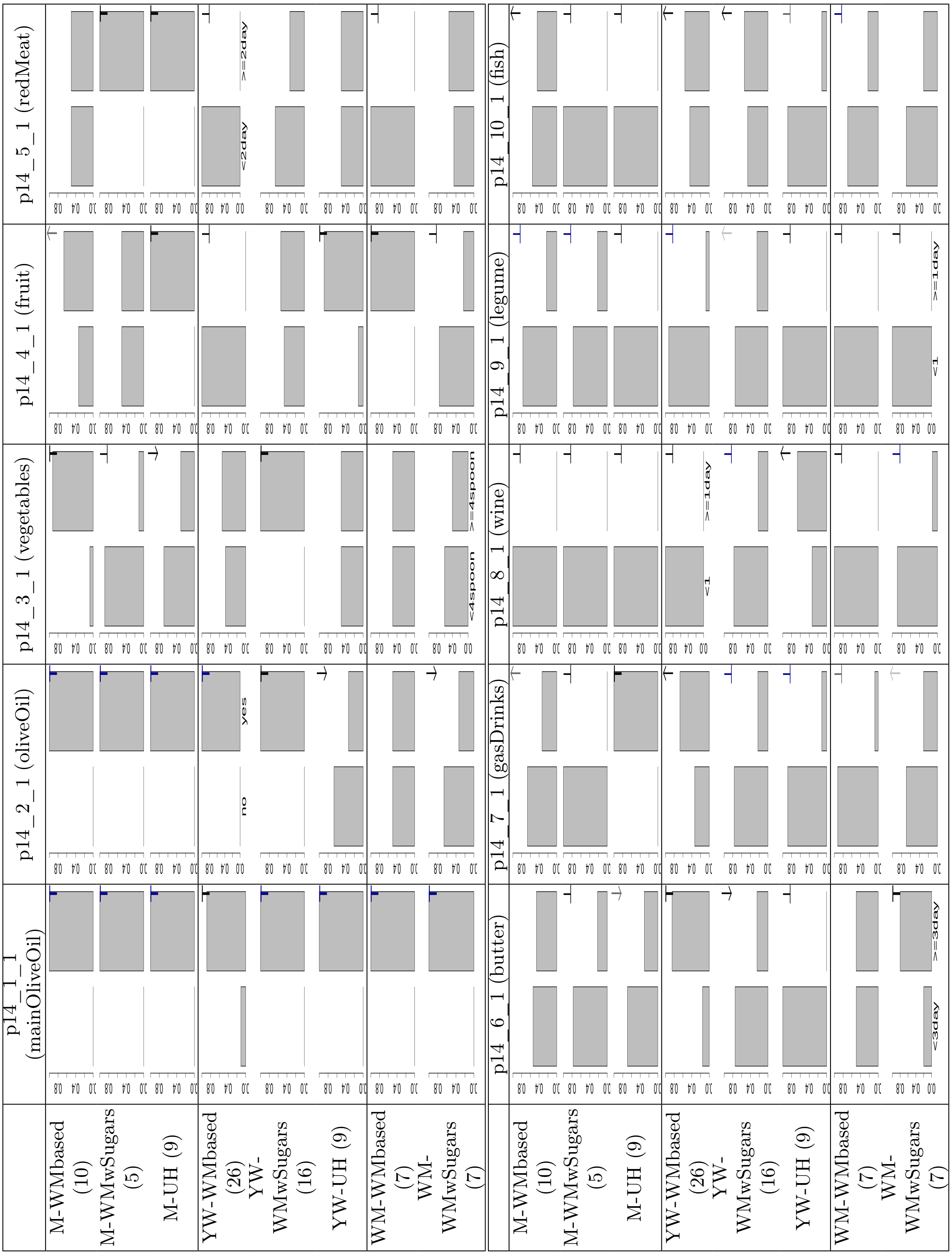


Table B.28: Diet characteristics ∼ Cross (Pre) (a)

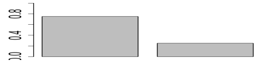
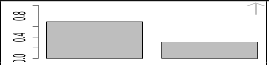
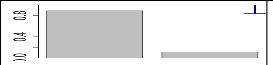
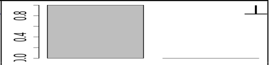
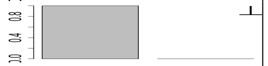
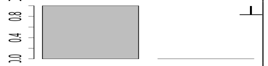
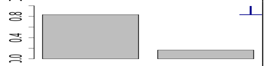
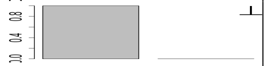
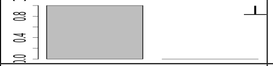
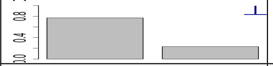
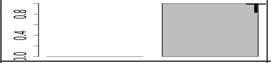
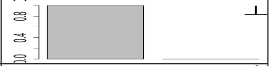
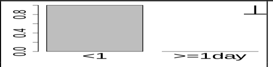
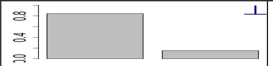
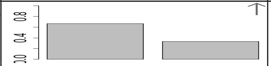
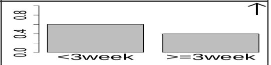
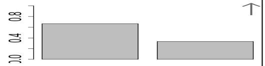
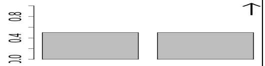
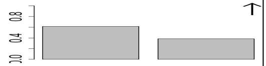
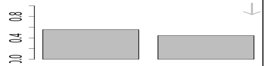
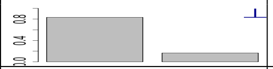
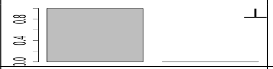
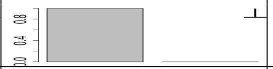
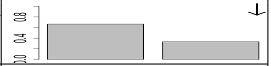
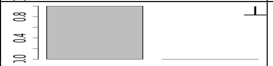
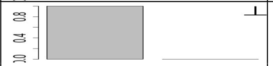
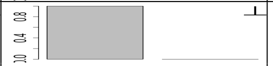
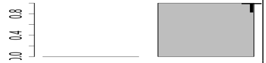
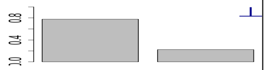
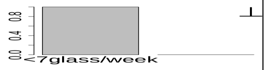
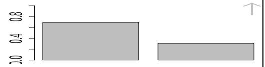
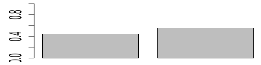
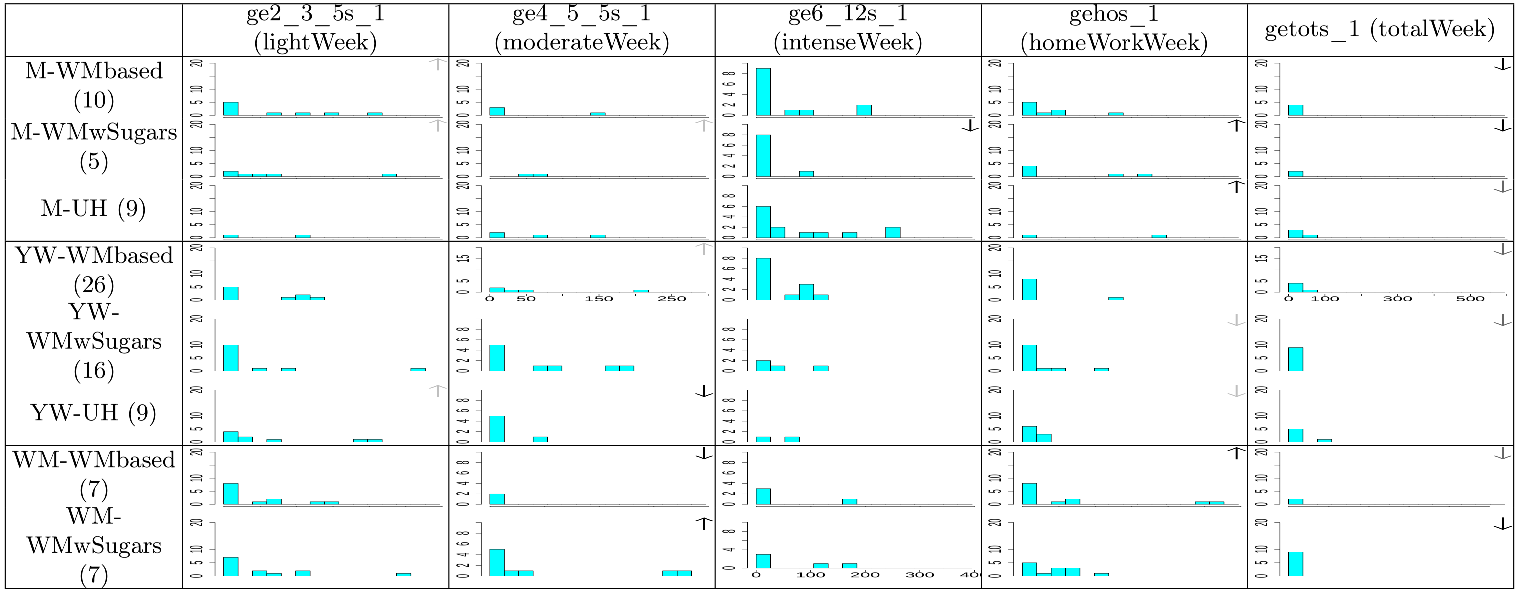
	p14_11_1 (commercialBakery)	p14_12_1 (nuts)	p14_13_1 (whiteMeat)	p14_14_1 (sauce)
M-WMbased (10)				
M-WMwSugars (5)				
M-UH (9)				
YW-WMbased (26)				
YW-WMwSugars (16)				
YW-UH (9)				
WM-WMbased (7)				
WM-WMwSugars (7)				

Table B.29: Diet characteristics ∼ Cross (Pre) (b)



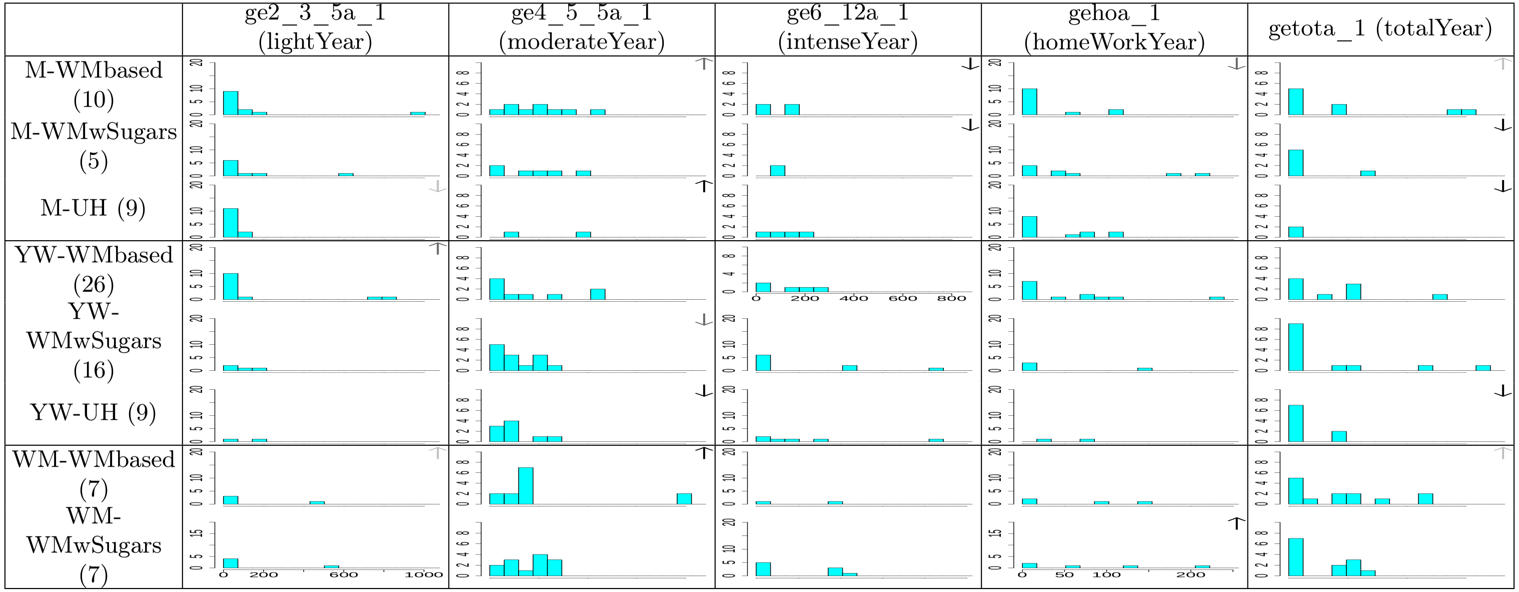


Table B.30: Physical activity characteristics ~ Cross (Pre)

B.3.3 Automatic Profiles of P_{H_o}

M-WMbased : **Less** Woman (-H) in *sexo* (gender); *altura1* (height) is higher(H); *peso1* (weight) is higher(H); *cint1* (waist) is higher(H); *pas_esds_2* (systolic Pressure) is higher(H); *pad_esds_2* (diastolic Pressure) is higher(HM); *fc_d_2* (heart Rate) is lower(-HM); More ex>5 () in *tabaco* (tobacco); Less soltero () and more Viudo/a (S) in *est_civi* (civilState); **Less** TRUE (-H) in *menopausia_bin* (menopause); *pertensa* (stressful) is lower(-H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea* (dyspnea); Less TRUE (-H) in *maspirina1* (aspirin); **Less** TRUE (-H) in *mdolor1* (NSAID); **Less** TRUE (-H) in *mansiedad1* (anxiolytic); **Less** TRUE (-H) in *mvitMin1* (vitamin or minerals); Less TRUE (-H) in *mtension1* (heart); Less TRUE (-H) in *mhormo1* (hormon); **Less** TRUE (-H) in *motros1* (other); *hdl0* (HDL) is lower(-H); *isoprostanos_0* (F2 α Isoprostanes) is higher(HM); *mcp_1_0* (Monocyte Chemotactic Protein-1) is higher(H); *il_6_0* (Interleukin 6) is lower(-SL); *il_10_0* (Interleukin 10) is lower(-SM); *tyru0* (Tyrosol) is higher(S); *ohtyru0* (OHTyrosol) is higher(S); *mohtyu0* (MOH_Tyrosol) is higher(S); More yes (H) in *p14_1_1* (mainOliveOil); **Less** >=3day (-H) in *p14_4_1* (fruit); Less >=1day (-H) in *p14_5_1* (redMeat); More >=1day (H) in *p14_6_1* (butter); Less >=1day (-H) in *p14_7_1* (gasDrinks); >=7glass/week (WI) in *p14_8_1* (wine); Less >=3week (-H) in *p14_9_1* (legume); **Less** >=2week (-H) in *p14_11_1* (commercialBakery); More yes (H) in *p14_13_1* (whiteMeat); More >=2week (H) in *p14_14_1* (sauce); *ge6_12s_1* (intenseWeek) is higher(S); *gehos_1* (homeWorkWeek) is lower(); *getots_1* (totalWeek) is higher(S); *ge6_12a_1* (intenseYear) is higher(S); *gehoa_1* (homeWorkYear) is lower(); *getota_1* (totalYear) is higher(S);

M-WMwSugars : **Less** Woman (-H) in *sexo* (gender); *altura1* (height) is higher(H); *peso1* (weight) is higher(H); *imc1* (BMI) is higher(HM); *cint1* (waist) is higher(H); *pas_esds_2* (systolic Pressure) is higher(H); *pad_esds_2* (diastolic Pressure) is higher(HW); More Never () in *tabaco* (tobacco); **Less** TRUE (-H) in *menopausia_bin* (menopause); *pertensa* (stressful) is lower(-H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); More TRUE (H) in *fractura* (bone fracture); More TRUE (H) in *disnea* (dyspnea); Less TRUE (-H) in *maspirina1* (aspirin); **Less** TRUE (-H) in *mdolor1* (NSAID); **Less** TRUE (-H) in *mansiedad1* (anxiolytic); **Less** TRUE (-H) in *mvitMin1* (vitamin or minerals); Less TRUE (-H) in *mtension1* (heart); Less TRUE (-H) in *mhormo1* (hormon); More TRUE (H) in *motros1* (other); *gluc0* (glucose) is higher(HW); *LDL_0* (LDL) is lower(-SL); *hdl0* (HDL) is lower(-H); *tryg0* (triglycerides) is higher(H); *mcp_1_0* (Monocyte Chemotactic Protein-1) is higher(H); *s_p_selectin_0* (sP-selectin) is higher(SW); *s_cd40l_0* (sCD40 Ligand) is lower(-S); *pcr0* (C-Reactive Protein) is lower(); *t_pa_0* (Tissue Plasminogen Activator) is higher(HW); *tyru0* (Tyrosol) is lower(-SM); More yes (H) in *p14_1_1* (mainOliveOil); **Less** >=2day (-H) in *p14_3_1* (vegetables); **More** >=3day (H) in *p14_4_1* (fruit); Less >=1day (-H) in *p14_5_1* (redMeat); **Less** >=1day (-H) in *p14_6_1* (butter); Less >=1day (-H) in *p14_7_1* (gasDrinks); More >=7glass/week (SW) in *p14_8_1* (wine); More >=3week (S) in *p14_9_1* (legume); **More** >=2week (H) in *p14_11_1* (commercialBakery); >=3week (WI) in *p14_12_1* (nuts); More yes (H) in *p14_13_1* (whiteMeat); More >=2week (H) in *p14_14_1* (sauce); *ge2_3_5s_1* (lightWeek) is higher(S); *ge4_5_5s_1* (moderateWeek) is higher(H); *ge6_12s_1* (intenseWeek) is lower(-S); *gehos_1* (homeWorkWeek) is lower(-S); *getots_1* (totalWeek) is higher(H); *ge2_3_5a_1* (lightYear) is higher(S); *ge4_5_5a_1* (moderateYear) is higher(H); *ge6_12a_1* (intenseYear) is lower(-SM); *gehoa_1* (homeWorkYear) is lower(-S); *getota_1* (totalYear) is higher(H);

M-UH : **Less** Woman (-H) in *sexo* (gender); *EDAD* (Age) is lower(-SW); *altura1* (height) is higher(H); *peso1* (weight) is higher(H); *imc1* (BMI) is higher(H); *cint1* (waist) is higher(H); *pas_esds_2* (systolic Pressure) is higher(HW); *fc_d_2* (heart Rate) is lower(-H); More ex0-1 (S) and more ex1-5 (S) and less Never (-S) in *tabaco* (tobacco); Less casado/a () and more soltero (S) in *est_civi* (civilState); **Less** TRUE (-H) in *menopausia_bin* (menopause); *pertensa* (stressful) is lower(-H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea* (dyspnea); Less TRUE (-H) in *maspirina1* (aspirin); **Less** TRUE (-H) in *mdolor1* (NSAID); **Less** TRUE (-H) in *mansiedad1* (anxiolytic); Less TRUE (-H) in *mvitMin1* (vitamin or minerals); Less TRUE (-H) in *mtension1* (heart); Less

TRUE (-H) in *mhormo1* (hormon); **Less** TRUE (-H) in *motros1* (other); *cholest0* (cholesterol) is lower(-HW); *LDL_0* (LDL) is lower(-SW); *hdl0* (HDL) is lower(-H); *tryg0* (triglycerides) is higher(H); *isoprostanos_0* (F2 α Isoprostanes) is higher(HL); *mcp_1_0* (Monocyte Chemotactic Protein-1) is (WI); *s_p_selectin_0* (sP-selectin) is higher(SL); *s_cd40l_0* (sCD40 Ligand) is higher(S); *oxo_gg_0* (8-Oxoguanine) is lower(-SM); *t_pa_0* (Tissue Plasminogen Activator) is higher(H); More yes (H) in *p14_1_1* (mainOliveOil); **Less** >=2day (-H) in *p14_3_1* (vegetables); **More** >=1day (H) in *p14_5_1* (redMeat); Less >=1day (-H) in *p14_6_1* (butter); More >=1day (H) in *p14_7_1* (gasDrinks); Less >=7glass/week (-H) in *p14_8_1* (wine); More >=3week (SL) in *p14_9_1* (legume); **Less** >=3week (-H) in *p14_10_1* (fish); **Less** >=3week (-H) in *p14_12_1* (nuts); More yes (H) in *p14_13_1* (whiteMeat); More >=2week (H) in *p14_14_1* (sauce); *ge4_5_5s_1* (moderateWeek) is lower(-HM); *gehos_1* (homeWorkWeek) is lower(-S); *getots_1* (totalWeek) is lower(-H); *ge2_3_5a_1* (lightYear) is lower(); *ge4_5_5a_1* (moderateYear) is lower(-H); *gehoa_1* (homeWorkYear) is lower(-S); *getota_1* (totalYear) is lower(-H);

YW-WMbased : **More** Woman (H) in *sexo* (gender); *EDAD* (Age) is lower(-HL); *peso1* (weight) is lower(-H); *imc1* (BMI) is lower(-HM); *cint1* (waist) is lower(-H); *pas_esds_2* (systolic Pressure) is lower(-H); *pad_esds_2* (diastolic Pressure) is lower(-HW); Less ex0-1 () and more Yes () in *tabaco* (tobacco); **Less** TRUE (-H) in *menopausia_bin* (menopause); *pertensa* (stressful) is higher(HW); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea* (dyspnea); Less TRUE (-H) in *maspirina1* (aspirin); More TRUE (H) in *mdolor1* (NSAID); Less TRUE (-H) in *mansiedad1* (anxiolytic); Less TRUE (-H) in *mvitMin1* (vitamin or minerals); **Less** TRUE (-H) in *mtension1* (heart); Less TRUE (-H) in *mhormo1* (hormon); *gluc0* (glucose) is lower(-HM); *hdl0* (HDL) is higher(HM); *tryg0* (triglycerides) is lower(-HW); *mcp_1_0* (Monocyte Chemotactic Protein-1) is lower(-SM); *pcr0* (C-Reactive Protein) is higher(SM); More yes (H) in *p14_1_1* (mainOliveOil); **More** >=2day (H) in *p14_3_1* (vegetables); **Less** >=3day (-H) in *p14_4_1* (fruit); Less >=1day (-H) in *p14_5_1* (redMeat); Less >=1day (-H) in *p14_6_1* (butter); Less >=7glass/week (-H) in *p14_8_1* (wine); Less >=3week (-H) in *p14_9_1* (legume); **More** >=3week (H) in *p14_10_1* (fish); Less >=2week (-HW) in *p14_11_1* (commercialBakery); More yes (H) in *p14_13_1* (whiteMeat); More >=2week (H) in *p14_14_1* (sauce);

YW-WMwSugars : **More** Woman (H) in *sexo* (gender); *altura1* (height) is lower(-H); *peso1* (weight) is lower(-H); *imc1* (BMI) is lower(-HW); *cint1* (waist) is lower(-H); *pas_esds_2* (systolic Pressure) is lower(-HM); *pad_esds_2* (diastolic Pressure) is (-WI); More ex1-5 () in *tabaco* (tobacco); **Less** TRUE (-H) in *menopausia_bin* (menopause); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea* (dyspnea); Less TRUE (-H) in *maspirina1* (aspirin); Less TRUE (-H) in *mdolor1* (NSAID); Less TRUE (-H) in *mansiedad1* (anxiolytic); Less TRUE (-H) in *mvitMin1* (vitamin or minerals); Less TRUE (-H) in *mtension1* (heart); Less TRUE (-H) in *mhormo1* (hormon); Less TRUE (-S) in *motros1* (other); *cholest0* (cholesterol) is lower(); *LDL_0* (LDL) is lower(); *tryg0* (triglycerides) is lower(-H); *oxldl0* (Oxidized LDL) is lower(); *ifn_g_0* (Interferon-γ) is higher(S); *pcr0* (C-Reactive Protein) is lower(-S); *tnf_a_0* (Tumor necrosis factor-α) is higher(SM); *t_pa_0* (Tissue Plasminogen Activator) is lower(-H); *ohtyru0* (OHTyrosol) is higher(); *mohtyu0* (MOH_Tyrosol) is higher(SM); More yes (H) in *p14_1_1* (mainOliveOil); More >=4spoon (HL) in *p14_2_1* (oliveOil); **More** >=3day (H) in *p14_4_1* (fruit); **Less** >=1day (-H) in *p14_5_1* (redMeat); Less >=1day (-H) in *p14_6_1* (butter); Less >=3week (-S) in *p14_9_1* (legume); **More** >=2week (H) in *p14_11_1* (commercialBakery); More >=3week (H) in *p14_12_1* (nuts); More yes (H) in *p14_13_1* (whiteMeat); More >=2week (H) in *p14_14_1* (sauce); *ge4_5_5s_1* (moderateWeek) is (WI); *ge6_12s_1* (intenseWeek) is higher(S); *getots_1* (totalWeek) is higher(H); *ge4_5_5a_1* (moderateYear) is (WI); *ge6_12a_1* (intenseYear) is higher(S); *getota_1* (totalYear) is higher(H);

YW-UH : **More** Woman (H) in *sexo* (gender); *EDAD* (Age) is lower(-HW); *altura1* (height) is lower(-H); *pas_esds_2* (systolic Pressure) is lower(-HW); *pad_esds_2* (diastolic Pressure) is lower(-H); More ex0-1 (S) in *tabaco* (tobacco); More Separado/a (SM) in *est_civi* (civilState); **Less** TRUE (-H) in *menopausia_bin* (menopause); *pertensa* (stressful) is higher(H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea* (dyspnea); Less TRUE (-H) in *maspirina1* (aspirin); **Less** TRUE (-H) in *mdolor1* (NSAID); Less TRUE (-H) in *mansiedad1* (anxiolytic); Less TRUE (-H) in *mvitMin1* (vitamin or minerals); Less TRUE (-H) in *mtension1* (heart); Less TRUE (-H) in *mhormo1* (hormon); More TRUE () in *motros1* (other); *gluc0* (glucose) is lower(-H); *cholest0* (cholesterol) is lower(-S); *LDL_ 0* (LDL) is lower(-S); *tryg0* (triglycerides) is lower(-H); *oxldl0* (Oxidized LDL) is lower(-HW); *isoprostanos_ 0* (F2 α Isoprostanes) is higher(); *oxo_ gg_ 0* (8-Oxoguanine) is higher(); *tnf_ a_ 0* (Tumor necrosis factor-α) is higher(SM); *t_ pa_ 0* (Tissue Plasminogen Activator) is lower(-H); *s_ vcam_ 1_ 0* (Soluble cell adhesion molecules-1) is higher(); *il_ 6_ 0* (Interleukin 6) is higher(S); *il_ 10_ 0* (Interleukin 10) is lower(-SM); *tyru0* (Tyrosol) is higher(S); More yes (H) in *p14_ 1_ 1* (mainOliveOil); **Less** >=4spoon (-HW) in *p14_ 2_ 1* (oliveOil); **Less** >=2day (-H) in *p14_ 3_ 1* (vegetables); More >=1day (HM) in *p14_ 5_ 1* (redMeat); More >=1day (H) in *p14_ 6_ 1* (butter); Less >=1day (-H) in *p14_ 7_ 1* (gasDrinks); Less >=7glass/week (-H) in *p14_ 8_ 1* (wine); Less >=3week (-SW) in *p14_ 9_ 1* (legume); **Less** >=3week (-H) in *p14_ 10_ 1* (fish); **More** >=2week (S) in *p14_ 11_ 1* (commercialBakery); **Less** >=3week (-H) in *p14_ 12_ 1* (nuts); More yes (H) in *p14_ 13_ 1* (whiteMeat); More >=2week (H) in *p14_ 14_ 1* (sauce); *ge2_ 3_ 5s_ 1* (lightWeek) is lower(); *ge4_ 5_ 5s_ 1* (moderateWeek) is lower(-HL); *ge6_ 12s_ 1* (intenseWeek) is lower(-HL); *gehos_ 1* (homeWorkWeek) is higher(S); *getots_ 1* (totalWeek) is lower(-H); *ge4_ 5_ 5a_ 1* (moderateYear) is lower(-HW); *ge6_ 12a_ 1* (intenseYear) is lower(-HL); *gehoa_ 1* (homeWorkYear) is higher(S); *getota_ 1* (totalYear) is lower(-H);

WM-WMbased : **More** Woman (H) in *sexo* (gender); *EDAD* (Age) is higher(H); *altura1* (height) is lower(-H); *peso1* (weight) is lower(-HW); *pad_ esds_ 2* (diastolic Pressure) is higher(H); *fc_ d_ 2* (heart Rate) is higher(H); More ex>5 (H) in *tabaco* (tobacco); More casado/a (SM) and more Divorciado/a (H) and less soltero (-H) in *est_ civi* (civilState); **More** TRUE (H) in *menopausia_ bin* (menopause); *pertensa* (stressful) is higher(SM); Less TRUE (-H) in *depre* (depression); More TRUE (H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea* (dyspnea); Less TRUE (-H) in *maspirina1* (aspirin); TRUE (WI) in *mdolor1* (NSAID); **Less** TRUE (-H) in *mansiedad1* (anxiolytic); Less TRUE (-H) in *mvitMin1* (vitamin or minerals); Less TRUE (-H) in *mtension1* (heart); Less TRUE (-H) in *mhormo1* (hormon); More TRUE (H) in *motros1* (other); *gluc0* (glucose) is higher(H); *cholest0* (cholesterol) is higher(H); *LDL_ 0* (LDL) is higher(H); *hdl0* (HDL) is higher(S); *oxldl0* (Oxidized LDL) is higher(H); *isoprostanos_ 0* (F2 α Isoprostanes) is lower(-H); *s_ p_ selectin_ 0* (sP-selectin) is lower(-H); *s_ cd40l_ 0* (sCD40 Ligand) is lower(-HM); *pcr0* (C-Reactive Protein) is higher(S); *il_ 8_ 0* (Interleukin 8) is higher(H); *il_ 10_ 0* (Interleukin 10) is higher(H); *tyru0* (Tyrosol) is lower(-HM); More yes (H) in *p14_ 1_ 1* (mainOliveOil); **More** >=2day (H) in *p14_ 3_ 1* (vegetables); **Less** >=3day (-H) in *p14_ 4_ 1* (fruit); **Less** >=1day (-H) in *p14_ 5_ 1* (redMeat); **Less** >=1day (-H) in *p14_ 6_ 1* (butter); **Less** >=1day (-H) in *p14_ 7_ 1* (gasDrinks); Less >=7glass/week (-H) in *p14_ 8_ 1* (wine); **Less** >=3week (-H) in *p14_ 9_ 1* (legume); >=3week (WI) in *p14_ 10_ 1* (fish); Less >=2week (-H) in *p14_ 11_ 1* (commercialBakery); Less >=3week (-S) in *p14_ 12_ 1* (nuts); More yes (H) in *p14_ 13_ 1* (whiteMeat); More >=2week (H) in *p14_ 14_ 1* (sauce); *ge2_ 3_ 5s_ 1* (lightWeek) is higher(S); *ge4_ 5_ 5s_ 1* (moderateWeek) is lower(); *ge6_ 12s_ 1* (intenseWeek) is lower(-S); *ge2_ 3_ 5a_ 1* (lightYear) is higher(S); *ge4_ 5_ 5a_ 1* (moderateYear) is lower(-SM); *ge6_ 12a_ 1* (intenseYear) is lower(-S); *getota_ 1* (totalYear) is lower(-SM);

WM-WMwSugars : **More** Woman (H) in *sexo* (gender); *EDAD* (Age) is higher(H); *altura1* (height) is lower(-H); *imc1* (BMI) is higher(S); *cint1* (waist) is higher(); *pas_ esds_ 2* (systolic Pressure) is higher(HM); *pad_ esds_ 2* (diastolic Pressure) is higher(H); *fc_ d_ 2* (heart Rate) is higher(H); More ex>5 (H) and more ex0-1 (S) and less Yes (-HM) in *tabaco* (tobacco); More Separado/a (S) and less soltero (-H) and more Viudo/a (HM) in *est_ civi* (civilState); **More** TRUE (H) in *menopausia_ bin* (menopause); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea* (dyspnea); Less TRUE (-H) in *maspirina1* (aspirin); More TRUE (H) in *mdolor1* (NSAID); **Less** TRUE (-H) in *mansiedad1* (anxiolytic); Less TRUE (-H) in *mvitMin1* (vitamin or minerals); Less TRUE (-H) in *mtension1* (heart); Less TRUE (-H) in *mhormo1* (hormon); **Less** TRUE (-H) in *motros1* (other); *gluc0* (glucose) is higher(HM); *cholest0* (cholesterol) is higher(H); *LDL_ 0* (LDL) is higher(H);

hdl0 (HDL) is lower(-SM); *tryg0* (triglycerides) is higher(HM); *oxldl0* (Oxidized LDL) is higher(H); *mcp_1_0* (Monocyte Chemotactic Protein-1) is lower(-HW); *s_p_selectin_0* (sP-selectin) is (-WI); *s_cd40l_0* (sCD40 Ligand) is lower(-H); *pcr0* (C-Reactive Protein) is lower(); *il_10_0* (Interleukin 10) is higher(H); *tyru0* (Tyrosol) is lower(-H); More yes (H) in *p14_1_1* (mainOliveOil); **Less** >=2day (-H) in *p14_3_1* (vegetables); **More** >=3day (H) in *p14_4_1* (fruit); **Less** >=1day (-H) in *p14_5_1* (redMeat); **Less** >=1day (-H) in *p14_6_1* (butter); **Less** >=1day (-H) in *p14_7_1* (gasDrinks); Less >=7glass/week (-SW) in *p14_8_1* (wine); **More** >=2week (H) in *p14_11_1* (commercialBakery); **More** >=3week (H) in *p14_12_1* (nuts); More yes (H) in *p14_13_1* (whiteMeat); More >=2week (H) in *p14_14_1* (sauce); *ge6_12s_1* (intenseWeek) is lower(-S); *gehos_1* (homeWorkWeek) is higher(S); *getots_1* (totalWeek) is (-WI); *ge4_5_5a_1* (moderateYear) is (WI); *ge6_12a_1* (intenseYear) is lower(); *gehoa_1* (homeWorkYear) is higher(S);

B.4 Addition Information about the Characterization of partition P_{C_f}

B.4.1 Descriptors of CI-IMS for P_{C_f}

	sexo (gender) - Woman	EDAD (Age)	altura2 (height)	peso2 (weight)	imc2 (BMI)	cintura2 (waist:2)	pas_esds_fin (Systolic pressure)	pad_esds_fin (Diastolic pressure)	fc_fin (Heart Rate)
M	↓**R**	↓M	↑R	↑R	↑R	↑R	↑R	↑R	↓R
W	↑**R**	↓W	↓R	↓R	↓R	↓R	↓R	↓R	↓R
WM	↑**R**	↑R	↓R	↓W	$\overline{\uparrow \mathrm{M}}$	$\overline{\uparrow \mathrm{R}}$	↑R	↑R	$\overline{\downarrow \mathrm{R}}$

Table B.31: Descriptors for post partition P_{C_f} of Biometrics characteristics

	tabaco2 (tobacco) - ex>5	tabaco2 (tobacco) - ex0-1	tabaco2 (tobacco) - ex1-5	tabaco2 (tobacco) - Never	tabaco2 (tobacco) - Yes
M	$\perp$ B	$\perp$ **W**	$\perp$ B	$\overline{\downarrow \text{R}}$	$\perp$ B
W	$\perp$ B	$\perp$ B	$\perp$ B	$\overline{\downarrow \text{R}}$	$\overline{\uparrow \text{R}}$
WM	$\perp$ B	$\perp$ B	$\perp$ **W**	$\overline{\uparrow \text{R}}$	$\downarrow$**R**

Table B.32: Descriptors for post partition P_{C_f} of Tobacco characteristics

	est_civ2 (civilState) - casado/a	est_civ2 (civilState) - Divorciado/a	est_civ2 (civilState) - Separado/a	est_civ2 (civilState) - soltero	est_civ2 (civilState) - Viudo/a
M	$\overline{\downarrow \text{R}}$	$\perp$ B	$\perp$ B	$\overline{\uparrow \text{R}}$	$\perp$ B
W	$\overline{\downarrow \text{R}}$	$\perp$ B	$\perp$ B	$\overline{\uparrow \text{R}}$	$\perp$ B
WM	$\top$ B	$\uparrow$R	$\overline{\uparrow \text{R}}$	$\downarrow$R	$\uparrow$W

Table B.33: Descriptors for post partition P_{C_f} of Sociodemographic characteristics

	menopausia_bin (menopause) - TRUE	pertensa (stressful)	depre (depression) - TRUE	cancer (cancer) - TRUE	fractura (bone fracture) - TRUE	disnea2 (dyspnea) - TRUE
M	**↓R**	↓R	⊥ B	**↓W**	⊥ B	⊥ B
W	**↓R**	↑R	⊥ B	⊥ B	⊥ B	⊥ B
WM	**↑R**	$\overline{\uparrow \mathrm{R}}$	⊥ B	↑R	⊥ B	⊥ B

Table B.34: Descriptors for post partition P_{C_f} of Menopause & Diseases characteristics

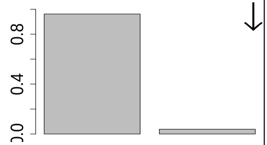
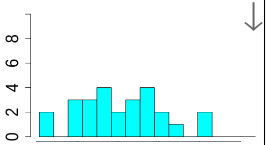
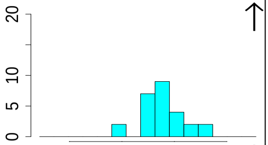
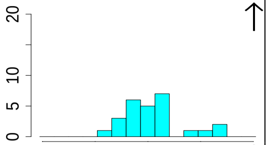
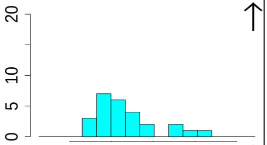
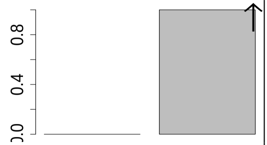
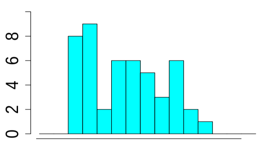
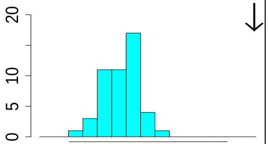
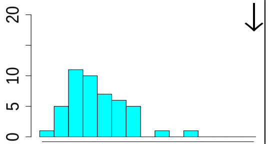
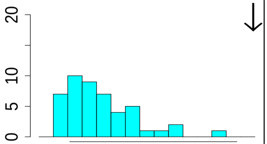
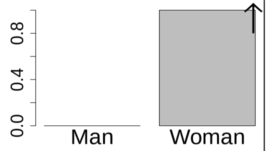
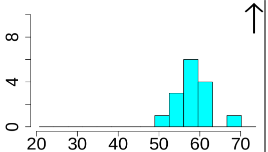
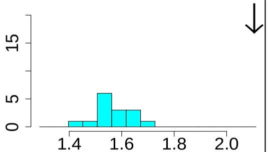
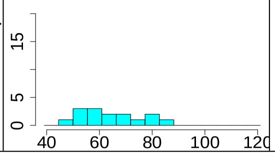
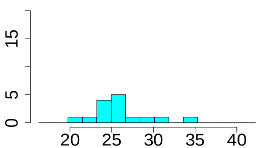
	maspirina2 (aspirin) - TRUE	mdolor2 (NSAID) - TRUE	mansiedad2 (anxiolytic) - TRUE	mvitMin2 (vitamin or minerals) - TRUE	mtension2 (heart) - TRUE	mhormo2 (hormon) - TRUE	motros2 (other) - TRUE
M	⊥ **R**	**↓R**	**↓M**	⊥ B	⊥ B	**↓W**	⊥ B
W	**↓W**	⊥ B	⊥ **R**	**↓W**	**↓M**	⊥ B	⊥ B
WM	⊥ B	$\overline{\uparrow \mathrm{R}}$	**↓M**	**↓W**	⊥ B	⊥ **M**	⊥ B

Table B.35: Descriptors for post partition P_{C_f} of Drugs characteristics

	gluc3 (glucose)	cholest1 (cholesterol)	LDL_3 (LDL)	hdl3 (HDL)	tryg3 (triglycerides)	oxldl3 (Oxidized LDL)	isoprostanos_3 (F2 α Isoprostanes)	ifn_g_3 (Interferon-γ)	mcp_1_3 (Monocyte Chemotactic Protein-1)	s_p_selectin_3 (sP-selectin)	s_cd40l_3 (sCD40 Ligand)	pcr3 (C-Reactive Protein)	oxo_gg_3 (8-Oxoguanine)	tnf_a_3 (Tumor necrosis factor-α)	t_pa_3 (Tissue Plasminogen Activator)	s_vcam_1_3 (Soluble cell adhesion molecules-1)	il_6_3 (Interleukin 6)	il_8_3 (Interleukin 8)	il_10_3 (Interleukin 10)	tyru3 (Tyrosol)	ohtyru3 (OHTyrosol)	mohtyu3 (MOH_Tyrosol)
M	$\overline{\downarrow \mathrm{R}}$	↓R	↓R	↓R	↑R	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{M}}$	↑R	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	↓R	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{W}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$
W	↓R	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	↓R	$\overline{\uparrow \mathrm{R}}$	↓R	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{W}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↓R	$\overline{\uparrow \mathrm{R}}$	↓R	↑W	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{M}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{W}}$
WM	$\overline{\uparrow \mathrm{W}}$	↑R	↑R	$\overline{\downarrow \mathrm{R}}$	↑R	↑R	↓M	$\overline{\downarrow \mathrm{R}}$	↓R	↓R	↓R	↑R	↓R	$\overline{\downarrow \mathrm{R}}$	↑R	$\overline{\downarrow \mathrm{W}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	↓R	$\overline{\downarrow \mathrm{W}}$	$\overline{\downarrow \mathrm{W}}$

Table B.36: Descriptors for post partition P_{C_f} of Biomarkers characteristics

B.4.2 Class Panel Graphs of partition P_{C_f}

	sexo (gender)	EDAD (Age)	altura2 (height)	peso2 (weight)	imc2 (BMI)
M (26)					
W (48)					
WM (15)					

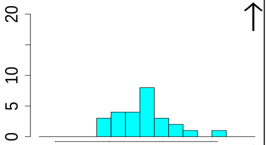
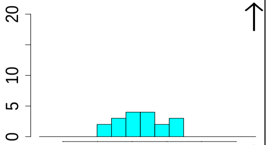
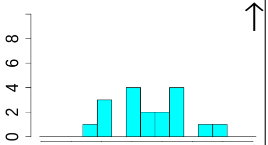
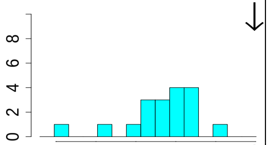
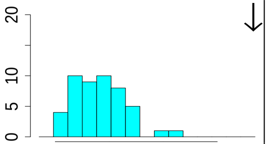
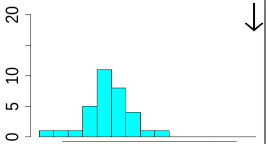
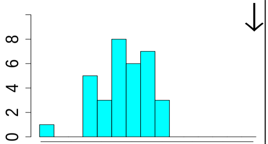
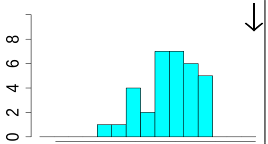
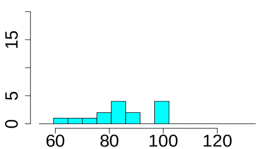
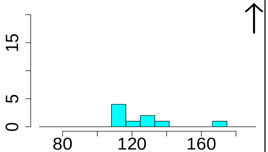
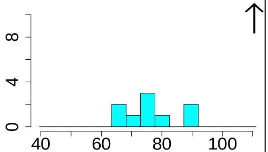
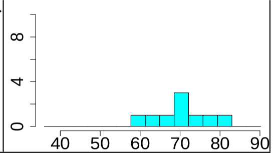
	cintura2 (waist:2)	pas_esds_fin (Systolic pressure)	pad_esds_fin (Diastolic pressure)	fc_fin (Heart Rate)
M (26)				
W (48)				
WM (15)				

Table B.37: Biometrics characteristics ∼ Baseline (Post)

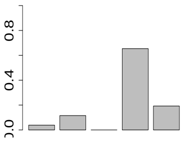
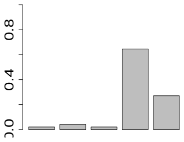
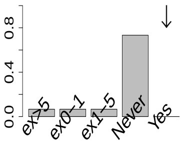
	tabaco2 (tobacco)
M (26)	
W (48)	
WM (15)	

Table B.38: Tobacco characteristics ∼ Baseline (Post)

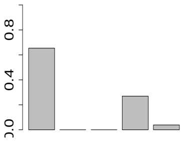
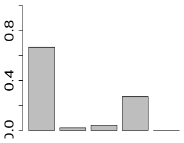
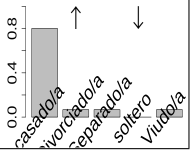
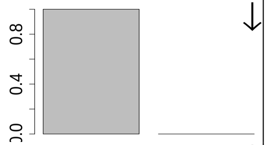
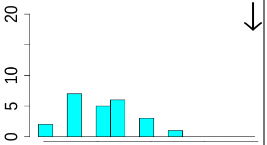
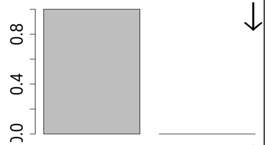
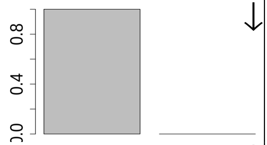
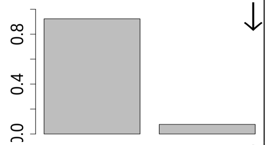
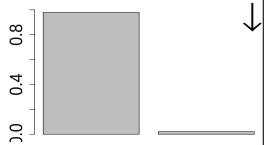
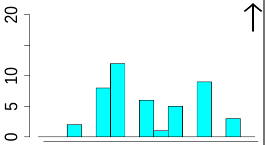
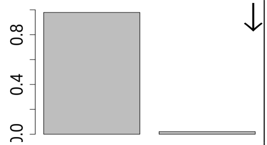
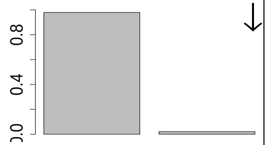
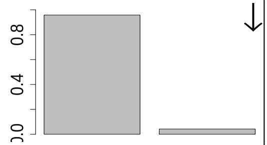
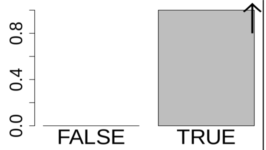
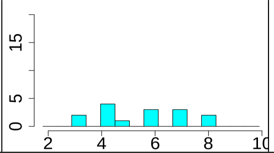
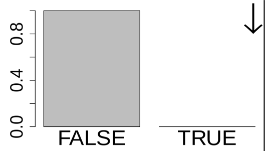
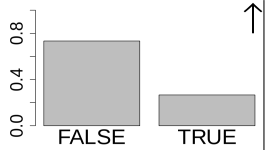
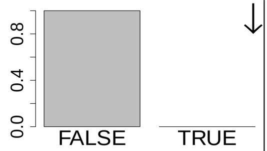
	est_civ2 (civilState)
M (26)	
W (48)	
WM (15)	

Table B.39: Sociodemographic characteristics ∼ Baseline (Post)

	menopausia_bin (menopause)	pertensa (stressful)	depre (depression)	cancer (cancer)	fractura (bone fracture)
M (26)					
W (48)					
WM (15)					

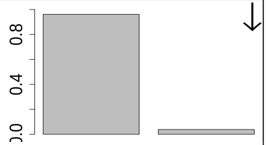
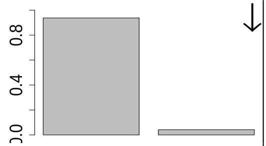
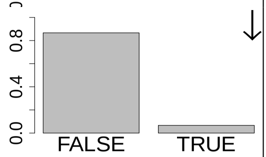
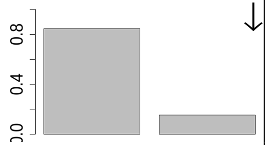
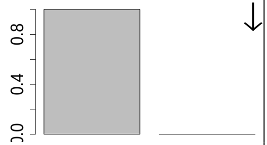
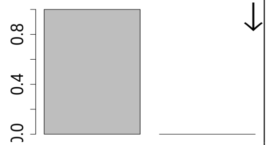
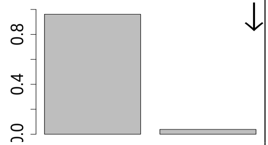
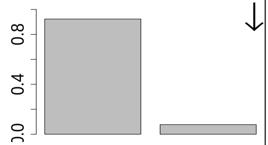
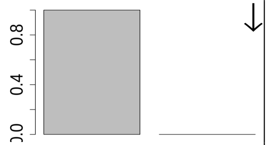
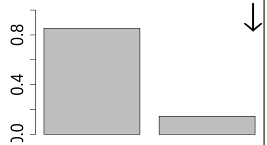
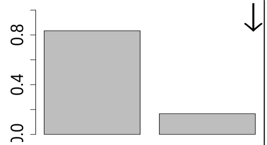
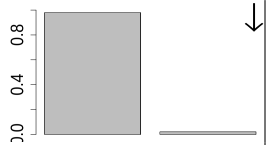
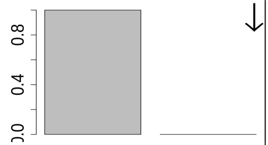
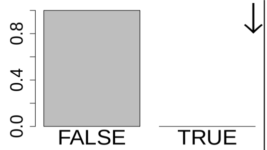
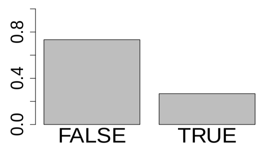
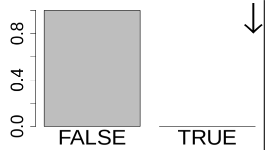
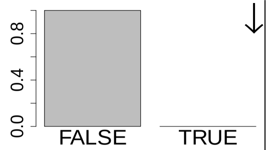
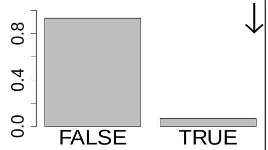
	disnea2 (dyspnea)
M (26)	
W (48)	
WM (15)	

Table B.40: Menopause & Diseases characteristics ~ Baseline (Post)

	maspirina2 (aspirin)	mdolor2 (NSAID)	mansiedad2 (anxiolytic)	mvitMin2 (vitamin or minerals)	mtension2 (heart)
M (26)					
W (48)					
WM (15)					

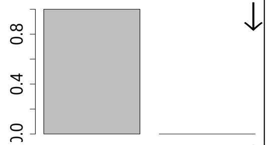
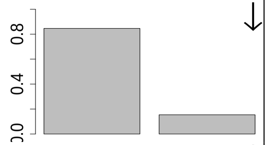
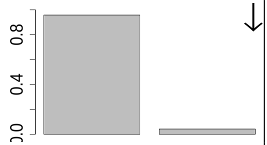
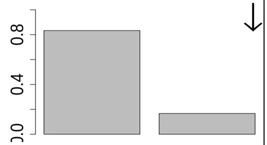
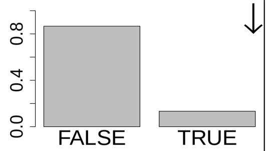
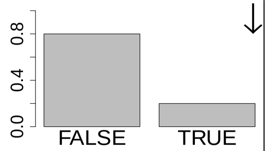
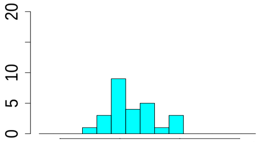
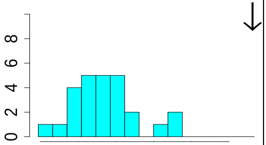
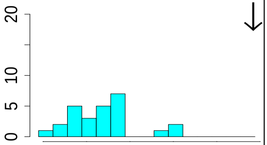
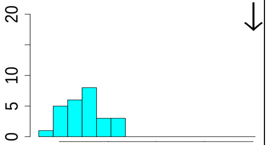
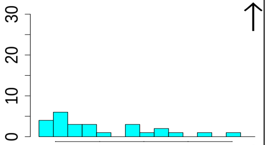
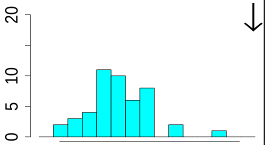
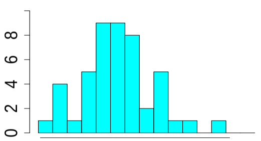
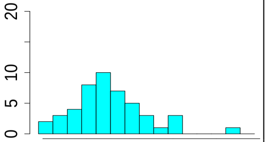
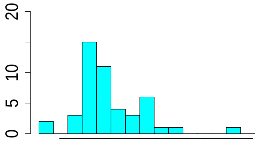
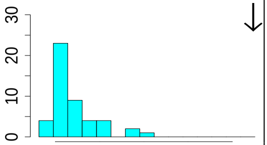
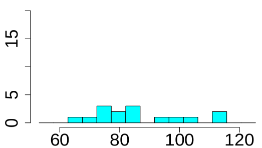
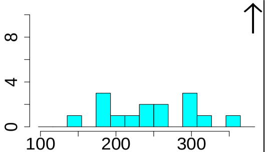
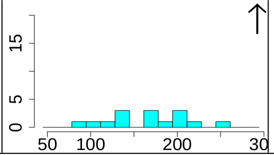
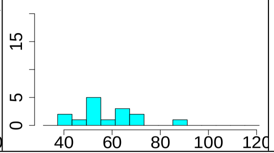
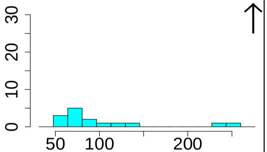
	mhormo2 (hormon)	motros2 (other)
M (26)		
W (48)		
WM (15)		

Table B.41: Drugs characteristics ~ Baseline (Post)

	gluc3 (glucose)	cholest1 (cholesterol)	LDL_3 (LDL)	hdl3 (HDL)	tryg3 (triglycerides)
M (26)					
W (48)					
WM (15)					

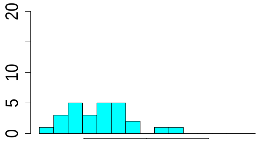
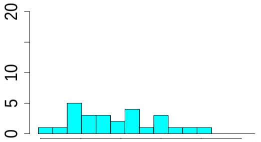
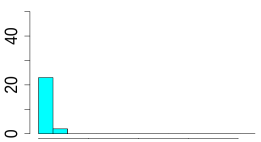
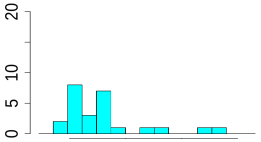
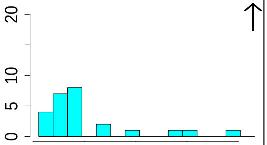
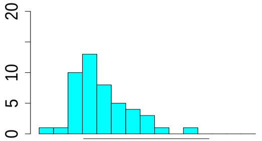
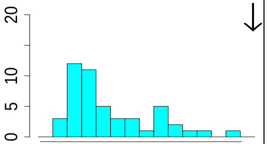
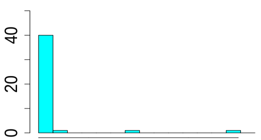
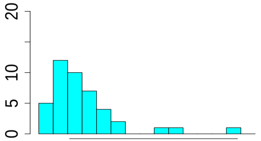
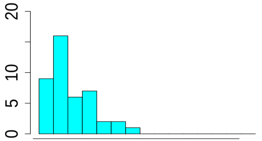
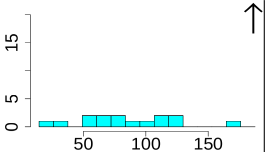
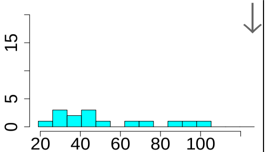
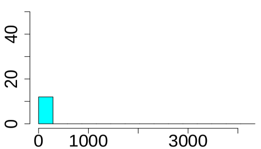
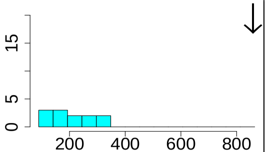
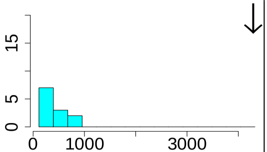
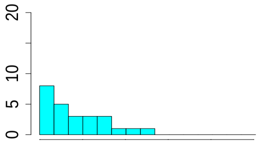
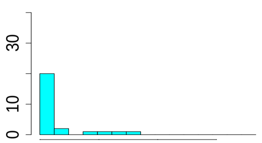
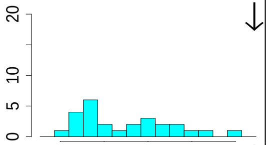
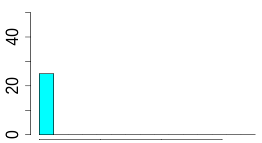
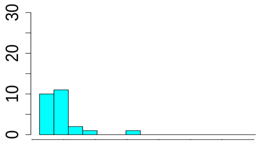
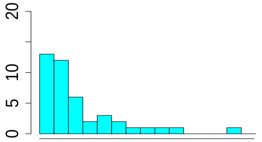
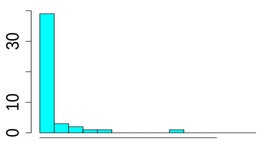
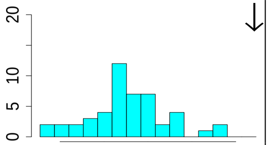
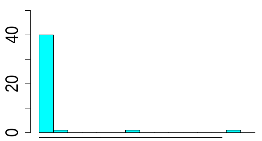
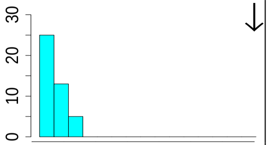
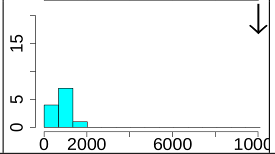
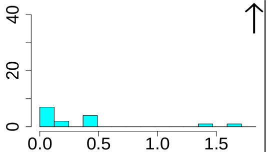
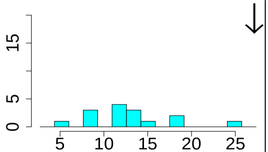
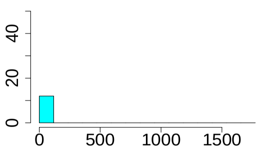
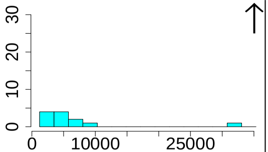
	oxldl3 (Oxidized LDL)	isoprostanos_3 (F2 α Isoprostanes)	ifn_g_3 (Interferon-γ)	mcp_1_3 (Monocyte Chemotactic Protein-1)	s_p_selectin_3 (sP-selectin)
M (26)					
W (48)					
WM (15)					

Table B.42: Biomarkers characteristics ∼ Baseline (Post) (a)

	s_cd40l_3 (sCD40 Ligand)	pcr3 (C-Reactive Protein)	oxo_gg_3 (8-Oxoguanine)	tnf_a_3 (Tumor necrosis factor-α)	t_pa_3 (Tissue Plasminogen Activator)
M (26)					
W (48)					
WM (15)					

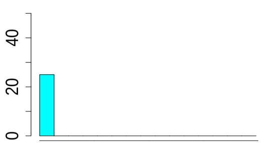
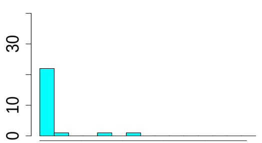
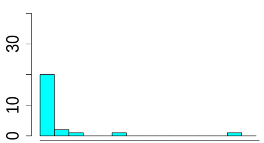
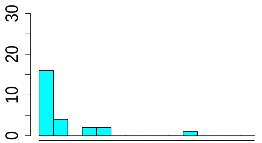
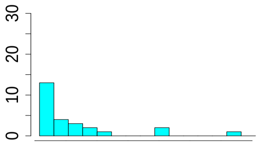
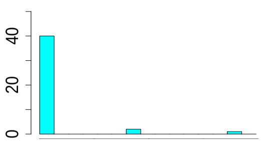
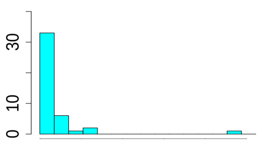
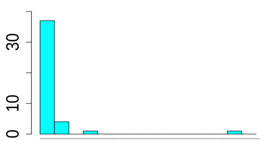
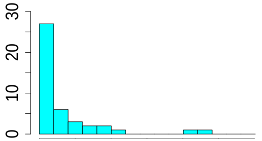
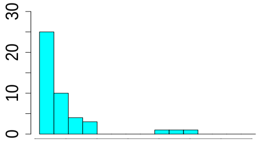
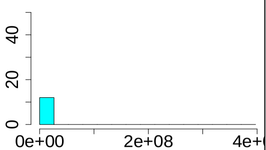
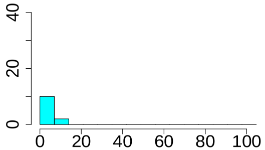
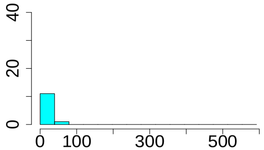
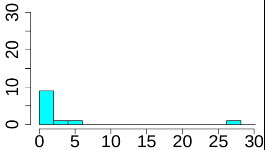
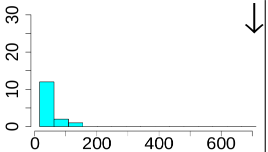
	s_vcam_1_3 (Soluble cell adhesion molecules-1)	il_6_3 (Interleukin 6)	il_8_3 (Interleukin 8)	il_10_3 (Interleukin 10)	tyru3 (Tyrosol)
M (26)					
W (48)					
WM (15)					

Table B.43: Biomarkers characteristics ~ Baseline (Post) (b)

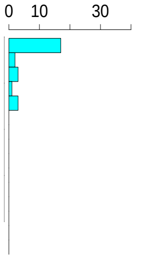
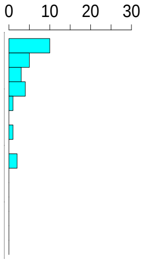
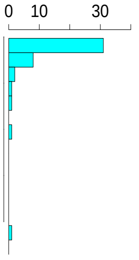
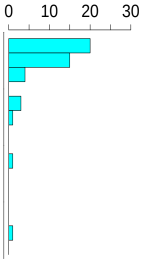
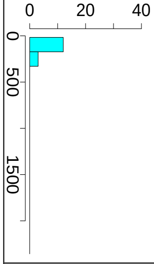
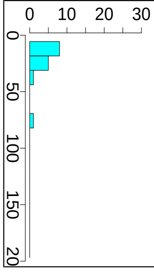
	ohtyru3 (OHTyrosol)	mohtyu3 (MOH_Tyrosol)
M (26)		
W (48)		
WM (15)		

Table B.44: Biomarkers characteristics ~ Baseline (Post) (c)

B.4.3 Automatic Profiles of P_{C_f}

M : Less Woman (-H) in *sexo* (gender); *EDAD* (Age) is lower(-HM); *altura2* (height) is higher(H); *peso2* (weight) is higher(H); *imc2* (BMI) is higher(H); *cintura2* (waist:2) is higher(H); *pas_ esds_ fin* (Systolic pressure) is higher(H); *pad_ esds_ fin* (Diastolic pressure) is higher(H); *fc_ fin* (Heart Rate) is lower(-H); More ex0-1 () in *tabaco2* (tobacco); Less TRUE (-H) in *menopausia_ bin* (menopause); *pertensa* (stressful) is lower(-H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); Less TRUE (-H) in *mdolor2* (NSAID); Less TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); Less TRUE (-H) in *motros2* (other); *cholest1* (cholesterol) is lower(-H); *LDL_ 3* (LDL) is lower(-H); *hdl3* (HDL) is lower(-H); *tryg3* (triglycerides) is higher(H); *s_ p_ selectin_ 3* (sP-selectin) is higher(H); *oxo_ gg_ 3* (8-Oxoguanine) is lower(-H);

W : More Woman (H) in *sexo* (gender); *EDAD* (Age) is lower(); *altura2* (height) is lower(-H); *peso2* (weight) is lower(-H); *imc2* (BMI) is lower(-H); *cintura2* (waist:2) is lower(-H); *pas_ esds_ fin* (Systolic pressure) is lower(-H); *pad_ esds_ fin* (Diastolic pressure) is lower(-H); *fc_ fin* (Heart Rate) is lower(-H); Less TRUE (-H) in *menopausia_ bin* (menopause); *pertensa* (stressful) is higher(H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); Less TRUE (-H) in *mdolor2* (NSAID); Less TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); Less TRUE (-H) in *motros2* (other); *gluc3* (glucose) is lower(-H); *tryg3* (triglycerides) is lower(-H); *isoprostanos_ 3* (F2 α Isoprostanes) is lower(-H); *oxo_ gg_ 3* (8-Oxoguanine) is lower(-H); *t_ pa_ 3* (Tissue Plasminogen Activator) is lower(-H); *s_ vcam_ 1_ 3* (Soluble cell adhesion molecules-1) is higher();

WM : More Woman (H) in *sexo* (gender); *EDAD* (Age) is higher(H); *altura2* (height) is lower(-H); *peso2* (weight) is lower(); *pas_ esds_ fin* (Systolic pressure) is higher(H); *pad_ esds_ fin* (Diastolic pressure) is higher(H); More ex1-5 () and less Yes (-H) in *tabaco2* (tobacco); More Divorciado/a (H) and less soltero (-H) and more Viudo/a () in *est_ civ2* (civilState); More TRUE (H) in *menopausia_ bin* (menopause); Less TRUE (-H) in *depre* (depression); More TRUE (H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); Less TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); Less TRUE (-H) in *motros2* (other); *cholest1* (cholesterol) is higher(H); *LDL_ 3* (LDL) is higher(H); *tryg3* (triglycerides) is higher(H); *oxldl3* (Oxidized LDL) is higher(H); *isoprostanos_ 3* (F2 α Isoprostanes) is lower(-HM); *mcp_ 1_ 3* (Monocyte Chemotactic Protein-1) is lower(-H); *s_ p_ selectin_ 3* (sP-selectin) is lower(-H); *s_ cd40l_ 3* (sCD40 Ligand) is lower(-H); *pcr3* (C-Reactive Protein) is higher(H); *oxo_ gg_ 3* (8-Oxoguanine) is lower(-H); *t_ pa_ 3* (Tissue Plasminogen Activator) is higher(H); *tyru3* (Tyrosol) is lower(-H);

B.5 Addition Information about the Characterization of partition P_{H_f}

B.5.1 Descriptors of CI-IMS for P_{H_f}

	p14_1_2 (mainOliveOil) - yes	p14_2_2 (oliveOil) - >=4spoon	p14_3_2 (vegetables) - >=2day	p14_4_2 (fruit) - >=3day	p14_5_2 (redMeat) - >=1day	p14_6_2 (butter) - >=1day	p14_7_2 (gasDrinks) - >=1day	p14_8_2 (wine) - >=7glass/week	p14_9_2 (legume) - >=3week	p14_10_2 (fish) - >=3week	p14_11_2 (commercialBakery) - >=2week	p14_12_2 (nuts) - >=3week	p14_13_2 (whiteMeat) - yes	p14_14_2 (sauce) - >=2week
HOO	$\top$ B	$\uparrow$**R**	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\perp$ B	$\perp$ **W**	$\uparrow$R	$\uparrow$R	$\perp$ B	$\uparrow$M	$\downarrow$R	$\downarrow$**R**	$\top$ B	$\top$ B
HB	$\top$ B	$\downarrow$**R**	$\uparrow$**R**	$\downarrow$M	$\downarrow$**R**	$\uparrow$R	$\downarrow$**R**	$\downarrow$**R**	$\downarrow$**M**	$\downarrow$R	$\downarrow$**R**	$\overline{\uparrow \mathrm{R}}$	$\top$ B	$\top$ B
BasicBak	$\top$ B	$\downarrow$M	$\uparrow$**R**	$\downarrow$M	$\perp$ B	$\downarrow$**R**	$\perp$ B	$\perp$ B	$\perp$ B	$\uparrow$W	$\uparrow$**R**	$\overline{\uparrow \mathrm{R}}$	$\top$ **M**	$\uparrow$**W**
ProtoCaloric	$\top$ B	$\overline{\downarrow \mathrm{R}}$	$\downarrow$**R**	$\uparrow$R	$\uparrow$R	$\perp$ B	$\overline{\uparrow \mathrm{R}}$	$\perp$ B	$\uparrow$R	$\overline{\uparrow \mathrm{R}}$	$\uparrow$**R**	$\uparrow$R	$\top$ B	$\top$ **R**

Table B.45: Descriptors for post partition P_{H_f} of Diet characteristics

	ge2_3_5s_2 (lightWeek)	ge4_5_5s_2 (moderateWeek)	ge6_12s_2 (intenseWeek)	gehos_2 (homeWorkWeek)	getots_2 (totalWeek)	ge2_3_5a_2 (lightYear)	ge4_5_5a_2 (moderateYear)	ge6_12a_2 (intenseYear)	gehoa_2 (homeWorkYear)	getota_2 (totalYear)
HOO	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$
HB	$\uparrow$M	$\downarrow$R	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{W}}$	$\downarrow$R	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{M}}$
BasicBak	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\uparrow$R	$\overline{\downarrow \mathrm{R}}$	$\uparrow$M	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$
ProtoCaloric	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$

Table B.46: Descriptors for post partition P_{H_f} of Physical activity characteristics

B.5.2 Class Panel Graphs of partition P_{H_f}

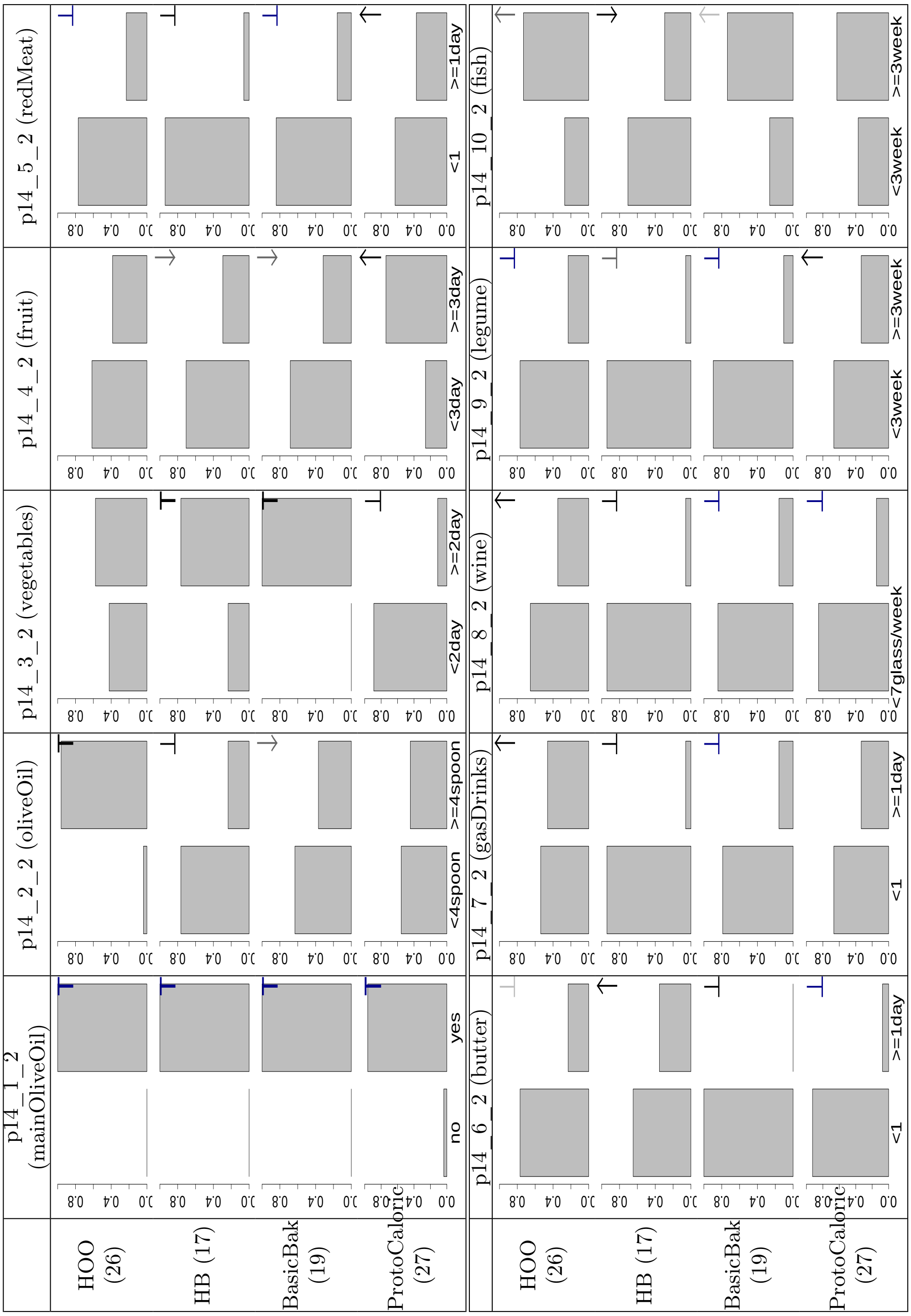


Table B.47: Diet characteristics ∼ Habits (Post) (a)

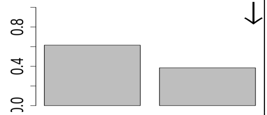
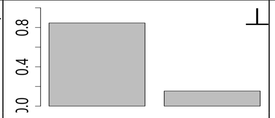
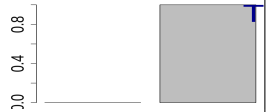
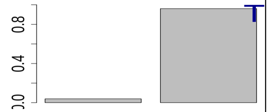
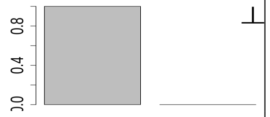
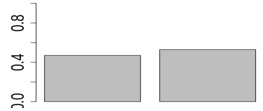
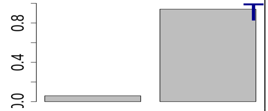
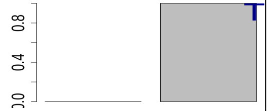
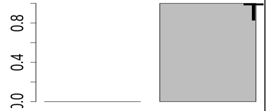
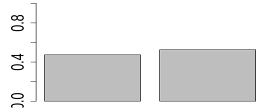
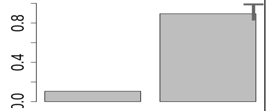
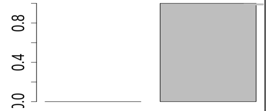
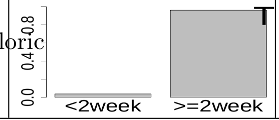
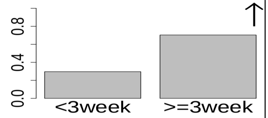
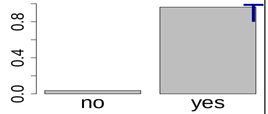
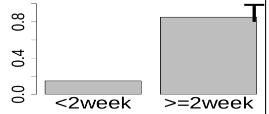
	p14_11_2 (commercialBakery)	p14_12_2 (nuts)	p14_13_2 (whiteMeat)	p14_14_2 (sauce)
HOO (26)				
HB (17)				
BasicBak (19)				
ProtoCaloric (27)				

Table B.48: Diet characteristics ∼ Habits (Post) (b)

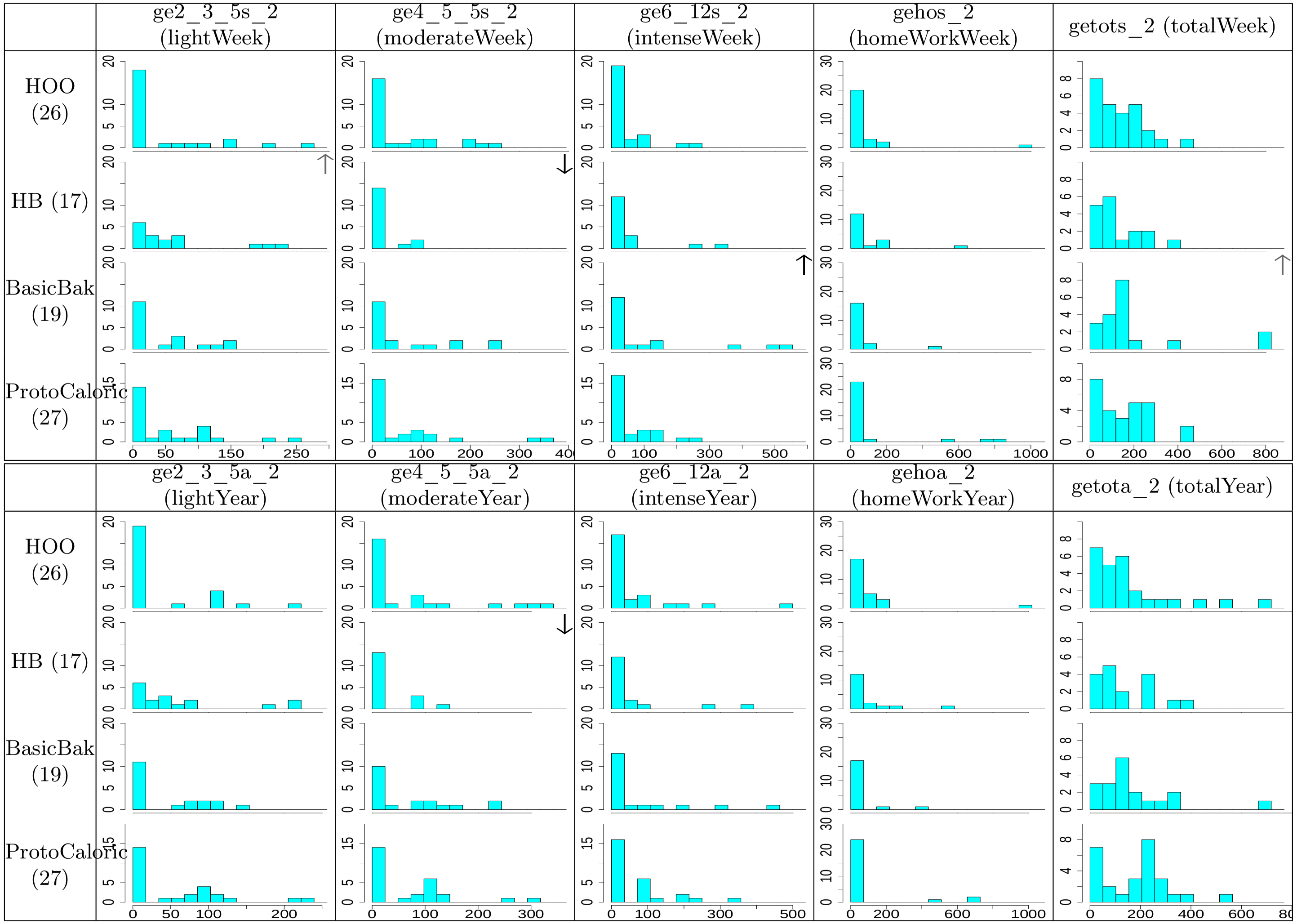


Table B.49: Physical activity characteristics ~ Habits (Post)

B.5.3 Automatic Profiles of P_{H_f}

H↑OO : More yes (H) in *p14_1_2* (mainOliveOil); More >=4spoon (H) in *p14_2_2* (oliveOil); Less >=1day (-H) in *p14_5_2* (redMeat); Less >=1day (-H) in *p14_6_2* (butter); More >=1day (H) in *p14_7_2* (gasDrinks); More >=7glass/week (H) in *p14_8_2* (wine); Less >=3week (-H) in *p14_9_2* (legume); More >=3week (HM) in *p14_10_2* (fish); Less >=2week (-H) in *p14_11_2* (commercialBakery); Less >=3week (-H) in *p14_12_2* (nuts); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce);

HB : More yes (H) in *p14_1_2* (mainOliveOil); Less >=4spoon (-H) in *p14_2_2* (oliveOil); More >=2day (H) in *p14_3_2* (vegetables); Less >=3day (-HM) in *p14_4_2* (fruit); Less >=1day (-H) in *p14_5_2* (redMeat); More >=1day (H) in *p14_6_2* (butter); Less >=1day (-H) in *p14_7_2* (gasDrinks); Less >=7glass/week (-H) in *p14_8_2* (wine); Less >=3week (-H) in *p14_9_2* (legume); Less >=3week (-H) in *p14_10_2* (fish); Less >=2week (-H) in *p14_11_2* (commercialBakery); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce); *ge2_3_5s_2* (lightWeek) is higher(HL); *ge4_5_5s_2* (moderateWeek) is lower(-H); *ge4_5_5a_2* (moderateYear) is lower(-H);

BasicBak : More yes (H) in *p14_1_2* (mainOliveOil); Less >=4spoon (-HM) in *p14_2_2* (oliveOil); More >=2day (H) in *p14_3_2* (vegetables); Less >=3day (-HM) in *p14_4_2* (fruit); Less >=1day (-H) in *p14_5_2* (redMeat); Less >=1day (-H) in *p14_6_2* (butter); Less >=1day (-H) in *p14_7_2* (gasDrinks); Less >=7glass/week (-H) in *p14_8_2* (wine); Less >=3week (-H) in *p14_9_2* (legume); More >=3week () in *p14_10_2* (fish); More >=2week (H) in *p14_11_2* (commercialBakery); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce); *ge6_12s_2* (intenseWeek) is higher(H); *getots_2* (totalWeek) is higher(HM);

ProtoCaloric : More yes (H) in *p14_1_2* (mainOliveOil); Less >=2day (-H) in *p14_3_2* (vegetables); More >=3day (H) in *p14_4_2* (fruit); More >=1day (H) in *p14_5_2* (redMeat); Less >=1day (-H) in *p14_6_2* (butter); Less >=7glass/week (-H) in *p14_8_2* (wine); More >=3week (H) in *p14_9_2* (legume); More >=2week (H) in *p14_11_2* (commercialBakery); More >=3week (H) in *p14_12_2* (nuts); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce);

B.6 Addition Information about the Characterization of partition P_f

B.6.1 Descriptors of CI-IMS for P_f

	sexo (gender) - Woman	EDAD (Age)	altura2 (height)	peso2 (weight)	imc2 (BMI)	cintura2 (waist:2)	pas_esds_fin (Systolic pressure)	pad_esds_fin (Diastolic pressure)	fc_fin (Heart Rate)
M-HOO	**↓R**	$\overline{\downarrow \mathrm{R}}$	↑R	↑R	↑W	↑R	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{W}}$	↓R
M-HB	**↓R**	↓M	↑R	↑R	↑R	↑R	↑R	↑M	↓R
M-BasicBak	**↓R**	$\overline{\uparrow \mathrm{R}}$	↑R	↑R	$\overline{\uparrow \mathrm{R}}$	↑R	↑R	↑R	↑W
M-ProtoCaloric	**↓R**	↓R	↑R	↑R	↑R	↑R	↑R	↑R	↓M
W-HOO	**↑R**	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{W}}$	↓R	↓W	↓W	↓R	↓R	↓M
W-HB	**↑R**	$\overline{\downarrow \mathrm{R}}$	↓R	↓R	$\overline{\downarrow \mathrm{R}}$	↓R	↓R	↓M	$\overline{\uparrow \mathrm{R}}$
W-BasicBak	**↑R**	$\overline{\downarrow \mathrm{R}}$	↓R	↓R	↓R	↓R	↓R	↓M	↓R
W-ProtoCaloric	**↑R**	↓R	↓W	↓R	↓M	↓R	↓R	$\overline{\downarrow \mathrm{R}}$	↓R
WM-HOO	**↑R**	↑R	↓R	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↑R	↑R	$\overline{\uparrow \mathrm{R}}$
WM-HB	**↑R**	↑R	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	↑M	↑R	↑R	↑R	↑W
WM-BasicBak	**↑R**	↑R	↓R	↓R	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↑M	$\overline{\uparrow \mathrm{R}}$	↓R
WM-ProtoCaloric	**↑R**	↑R	↓R	$\overline{\downarrow \mathrm{R}}$	↑M	↑R	↑R	↑R	$\overline{\downarrow \mathrm{R}}$

Table B.50: Descriptors for post partition P_f of Biometrics characteristics

	tabaco2 (tobacco) - ex>5	tabaco2 (tobacco) - ex0-1	tabaco2 (tobacco) - ex1-5	tabaco2 (tobacco) - Never	tabaco2 (tobacco) - Yes
M-HOO	⊥ **M**	⊥ **R**	⊥ B	$\overline{\downarrow \text{M}}$	⊥ B
M-HB	⊥ B	⊥ B	⊥ B	$\overline{\uparrow \text{R}}$	↑W
M-BasicBak	$\overline{\downarrow \text{R}}$	$\overline{\downarrow \text{W}}$	$\overline{\downarrow \text{R}}$	**↑R**	↓R
M-ProtoCaloric	⊥ B	⊥ B	⊥ B	$\overline{\uparrow \text{R}}$	⊥ B
W-HOO	⊥ B	⊥ **R**	⊥ **R**	↓W	⊥ B
W-HB	⊥ B	⊥ B	⊥ B	$\overline{\downarrow \text{M}}$	↑R
W-BasicBak	⊥ B	⊥ B	⊥ B	$\overline{\uparrow \text{R}}$	$\overline{\uparrow \text{W}}$
W-ProtoCaloric	$\overline{\downarrow \text{M}}$	$\overline{\downarrow \text{W}}$	$\overline{\downarrow \text{R}}$	⊤ B	$\overline{\uparrow \text{R}}$
WM-HOO	↑R	⊥ B	⊥ B	$\overline{\uparrow \text{R}}$	**↓R**
WM-HB	⊥ B	⊥ B	↑R	↓M	**↓R**
WM-BasicBak	$\overline{\downarrow \text{R}}$	$\overline{\downarrow \text{W}}$	$\overline{\downarrow \text{R}}$	**↑R**	↓R
WM-ProtoCaloric	⊥ B	⊥ **R**	⊥ B	$\overline{\downarrow \text{R}}$	**↓R**

Table B.51: Descriptors for post partition P_f of Tobacco characteristics

	est_civ2 (civilState) - casado/a	est_civ2 (civilState) - Divorciado/a	est_civ2 (civilState) - Separado/a	est_civ2 (civilState) - soltero	est_civ2 (civilState) - Viudo/a
M-HOO	↓W	⊥ B	⊥ B	⊥ B	⊥ **R**
M-HB	**↑W**	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{M}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$
M-BasicBak	**↑R**	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↓R	$\overline{\downarrow \mathrm{R}}$
M-ProtoCaloric	↓W	⊥ B	⊥ B	↑R	⊥ B
W-HOO	$\overline{\downarrow \mathrm{R}}$	⊥ **R**	⊥ B	⊥ B	⊥ B
W-HB	**↑R**	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{M}}$	↓R	$\overline{\downarrow \mathrm{R}}$
W-BasicBak	↓R	⊥ B	⊥ B	↑R	⊥ B
W-ProtoCaloric	$\overline{\downarrow \mathrm{R}}$	⊥ B	⊥ B	⊥ B	⊥ B
WM-HOO	$\overline{\uparrow \mathrm{R}}$	↑R	⊥ B	**↓R**	⊥ B
WM-HB	**↑R**	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↓R	$\overline{\downarrow \mathrm{R}}$
WM-BasicBak	**↑R**	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↓R	$\overline{\downarrow \mathrm{R}}$
WM-ProtoCaloric	$\overline{\downarrow \mathrm{R}}$	⊥ B	⊥ **R**	**↓R**	⊥ **R**

Table B.52: Descriptors for post partition P_f of Sociodemographic characteristics

	menopausia_bin (menopause) - TRUE	pertensa (stressful)	depre (depression) - TRUE	cancer (cancer) - TRUE	fractura (bone fracture) - TRUE	disnea2 (dyspnea) - TRUE
M-HOO	**↓R**	↓R	⊥ B	⊥ B	⊥ B	⊥ B
M-HB	**↓R**	↓R	⊥ B	⊥ B	⊥ B	⊥ B
M-BasicBak	**↓R**	$\overline{\downarrow \text{R}}$	⊥ B	⊥ B	⊥ B	⊥ B
M-ProtoCaloric	**↓R**	↓R	⊥ B	⊥ B	⊥ **R**	⊥ B
W-HOO	**↓R**	↑R	⊥ B	⊥ B	⊥ B	⊥ B
W-HB	**↓R**	↑W	⊥ B	⊥ B	⊥ **M**	⊥ **R**
W-BasicBak	**↓R**	↑W	⊥ **R**	⊥ B	⊥ B	⊥ B
W-ProtoCaloric	**↓W**	$\overline{\uparrow \text{M}}$	⊥ B	⊥ B	⊥ B	⊥ B
WM-HOO	**↑R**	$\overline{\downarrow \text{R}}$	⊥ B	↑R	⊥ B	⊥ B
WM-HB	**↑R**	↑R	⊥ B	↑R	⊥ B	⊥ B
WM-BasicBak	**↑R**	$\overline{\uparrow \text{R}}$	⊥ B	↑R	⊥ B	↑R
WM-ProtoCaloric	**↑R**	↓W	⊥ B	⊥ **R**	⊥ B	⊥ B

Table B.53: Descriptors for post partition P_f of Menopause & Diseases characteristics

	maspirina2 (aspirin) - TRUE	mdolor2 (NSAID) - TRUE	mansiedad2 (anxiolytic) - TRUE	mvitMin2 (vitamin or minerals) - TRUE	mtension2 (heart) - TRUE	mhormo2 (hormon) - TRUE	motros2 (other) - TRUE
M-HOO	⊥ **R**	**↓R**	**↓M**	⊥ B	⊥ B	⊥ B	**↓R**
M-HB	⊥ **R**	**↓R**	**↓W**	**↓W**	⊥ B	⊥ B	**↓R**
M-BasicBak	⊥ B	**↓R**	**↓W**	⊥ B	⊥ B	⊥ B	**↑R**
M-ProtoCaloric	⊥ **M**	**↓R**	**↓M**	**↓W**	⊥ **R**	⊥ B	⊥ B
W-HOO	⊥ B	**↓W**	⊥ B	**↓W**	⊥ B	⊥ B	**↓R**
W-HB	⊥ B	↑M	↑R	**↓W**	⊥ B	⊥ B	↑R
W-BasicBak	⊥ B	⊥ B	**↓M**	⊥ B	⊥ B	⊥ B	⊥ B
W-ProtoCaloric	⊥ B	**↓W**	↑R	**↓W**	⊥ B	⊥ B	**↓M**
WM-HOO	⊥ B	$\overline{\uparrow \text{R}}$	**↓W**	**↓W**	⊥ B	↑R	$\overline{\uparrow \text{R}}$
WM-HB	⊥ B	**↓R**	**↓W**	⊥ B	↑R	⊥ B	**↓R**
WM-BasicBak	⊥ B	↑R	**↓W**	**↓W**	⊥ B	⊥ B	**↓R**
WM-ProtoCaloric	⊥ B	⊥ B	**↓W**	**↓W**	⊥ B	⊥ **R**	↑R

Table B.54: Descriptors for post partition P_f of Drugs characteristics

	gluc3 (glucose)	cholest1 (cholesterol)	LDL_3 (LDL)	hdl3 (HDL)	tryg3 (triglycerides)	oxldl3 (Oxidized LDL)	isoprostanos_3 (F2 α Isoprostanes)	ifn_g_3 (Interferon-γ)	mcp_1_3 (Monocyte Chemotactic Protein-1)	s_p_selectin_3 (sP-selectin)	s_cd40l_3 (sCD40 Ligand)	pcr3 (C-Reactive Protein)	oxo_gg_3 (8-Oxoguanine)	tnf_a_3 (Tumor necrosis factor-α)	t_pa_3 (Tissue Plasminogen Activator)	s_vcam_1_3 (Soluble cell adhesion molecules-1)	il_6_3 (Interleukin 6)	il_8_3 (Interleukin 8)	il_10_3 (Interleukin 10)	tyru3 (Tyrosol)	ohtyru3 (OHTyrosol)	mohtyu3 (MOH_Tyrosol)
M-HOO	↓R	↑R	↑R	↓R	↑R	↑R	↓M	↓R	↑R	↑R	↓R	↑R	↓R	↓R	↑R	↓M	↑R	↓R	↑R	↑R	↓R	↑R
M-HB	↓R	↓R	↓R	↓W	↓R	↓R	↑R	↓R	↓R	↓W	↓R	↓W	↓R	↓R	↓M	↓M	↓R	↑W	↑R	↑R	↓R	↓R
M-BasicBak	↑R	↓R	↓R	↓R	↓W	↓R	↓R	↓R	↑R	↑R	↑W	↓M	↓R	↓R	↑R	↓M	↓M	↓R	↑R	↓R	↓W	↓W
M-ProtoCaloric	↑R	↓R	↓R	↓R	↑R	↓R	↓W	↓R	↑R	↑R	↓R	↓R	↓R	↓R	↑R	↓M	↓R	↓R	↓R	↓R	↓W	↓W
W-HOO	↓R	↑R	↑R	↑R	↓M	↓R	↓R	↓R	↓R	↓R	↓M	↓M	↓R	↓R	↓R	↑R	↓R	↑R	↑R	↓M	↓R	↓W
W-HB	↓W	↑R	↑R	↓R	↑R	↑R	↓R	↑R	↑R	↓R	↓W	↑R	↓R	↑R	↓W	↑R	↑R	↓R	↑R	↑R	↑R	↓R
W-BasicBak	↓R	↓R	↓R	↑R	↓R	↑R	↓R	↓R	↓M	↓M	↓R	↓R	↓R	↓R	↓R	↓W	↓M	↓R	↓W	↓R	↓W	↓R
W-ProtoCaloric	↓R	↓R	↓R	↓R	↓M	↓W	↓R	↑W	↑R	↑R	↑R	↓W	↓R	↑R	↓R	↓W	↓R	↓R	↓R	↓R	↓R	↓R
WM-HOO	↑R	↓R	↑R	↓R	↑R	↑R	↑W	↓R	↑R	↓R	↓R	↑R	↓R	↓R	↑M	↓M	↓M	↓R	↓W	↓R	↓R	↓M
WM-HB	↑R	↑R	↑R	↑R	↑M	↑R	↓R	↓R	↓R	↓R	↓R	↑R	↓R	↓R	↑R	↓M	↓M	↓R	↓R	↓R	↓W	↓W
WM-BasicBak	↑R	↑R	↑R	↑R	↑R	↑R	↓R	↓R	↓R	↓R	↓R	↑R	↓R	↓R	↓M	↓M	↑W	↓R	↑R	↓R	↓W	↓W
WM-ProtoCaloric	↓W	↑R	↑R	↓W	↑M	↑R	↓R	↓R	↓R	↓R	↓R	↓R	↓R	↓R	↑R	↓M	↓R	↓R	↓R	↓M	↓W	↓R

Table B.55: Descriptors for post partition P_f of Biomarkers characteristics

	p14_1_2 (mainOliveOil) - yes	p14_2_2 (oliveOil) - >=4spoon	p14_3_2 (vegetables) - >=2day	p14_4_2 (fruit) - >=3day	p14_5_2 (redMeat) - >=1day	p14_6_2 (butter) - >=1day	p14_7_2 (gasDrinks) - >=1day	p14_8_2 (wine) - >=7glass/week	p14_9_2 (legume) - >=3week	p14_10_2 (fish) - >=3week	p14_11_2 (commercialBakery) - >=2week	p14_12_2 (nuts) - >=3week	p14_13_2 (whiteMeat) - yes	p14_14_2 (sauce) - >=2week
M-HOO	⊤ B	↑**R**	↑M	$\overline{\downarrow \text{W}}$	↑M	⊥ B	↑R	↑M	⊥ B	↓W	↓**R**	↓**R**	⊤ B	⊤ B
M-HB	⊤ B	↓R	$\overline{\downarrow \text{R}}$	$\overline{\uparrow \text{R}}$	↓**R**	↑R	↓**R**	⊥ B	⊥ B	↓R	↓**R**	$\overline{\uparrow \text{R}}$	⊤ **R**	⊤ B
M-BasicBak	⊤ B	$\overline{\downarrow \text{R}}$	↑**R**	↓**R**	↑**R**	↓**R**	↓**R**	↓**R**	↑**R**	↑**R**	↑**R**	↑**R**	↓R	⊤ B
M-ProtoCaloric	⊤ **R**	↓R	↓**R**	$\overline{\uparrow \text{R}}$	↑R	⊥ B	↑R	⊥ B	↑M	$\overline{\downarrow \text{R}}$	↑**R**	$\overline{\downarrow \text{R}}$	⊤ **M**	⊤ B
W-HOO	⊤ B	↑**R**	$\overline{\uparrow \text{R}}$	$\overline{\uparrow \text{R}}$	⊥ B	⊥ B	↑R	↑W	↑R	↑**R**	$\overline{\downarrow \text{W}}$	↓**R**	⊤ B	⊤ B
W-HB	⊤ B	↓**R**	↑**R**	↓**R**	⊥ B	⊥ B	↓**M**	↓**R**	↓**R**	↓R	↓**R**	$\overline{\uparrow \text{R}}$	⊤ B	⊤ B
W-BasicBak	⊤ B	↓R	↑**R**	↓M	↓**M**	↓**R**	⊥ B	⊥ B	↓**R**	$\overline{\uparrow \text{R}}$	↑**R**	$\overline{\uparrow \text{R}}$	⊤ B	⊤ B
W-ProtoCaloric	⊤ B	$\overline{\uparrow \text{R}}$	↓**R**	↑**R**	↑W	⊥ B	$\overline{\uparrow \text{R}}$	⊥ B	↑W	$\overline{\downarrow \text{R}}$	↑**R**	↑**R**	⊤ B	⊤ **R**
WM-HOO	⊤ B	↑**R**	$\overline{\downarrow \text{R}}$	↓R	↓**R**	↑W	$\overline{\downarrow \text{R}}$	↑R	↓**R**	↑**R**	↓R	↓R	⊤ B	⊤ B
WM-HB	⊤ B	$\overline{\downarrow \text{R}}$	↑**R**	↓**R**	↓**R**	↓**R**	↓**R**	↓**R**	↓**R**	↓**R**	↓**R**	$\overline{\uparrow \text{R}}$	⊤ B	⊤ B
WM-BasicBak	⊤ B	$\overline{\downarrow \text{R}}$	↑**R**	$\overline{\uparrow \text{R}}$	↓**R**	↓**R**	↓**R**	↓**R**	↓**R**	↑W	↑**R**	↓R	⊤ B	⊤ B
WM-ProtoCaloric	⊤ B	$\overline{\downarrow \text{W}}$	↓**R**	↑**R**	↓**R**	↓**R**	↓**R**	↓**R**	↑R	↑**R**	↑**R**	↑**R**	⊤ B	⊤ **R**

Table B.56: Descriptors for post partition P_f of Diet characteristics

	ge2_3_5s_2 (lightWeek)	ge4_5_5s_2 (moderateWeek)	ge6_12s_2 (intenseWeek)	gehos_2 (homeWorkWeek)	getots_2 (totalWeek)	ge2_3_5a_2 (lightYear)	ge4_5_5a_2 (moderateYear)	ge6_12a_2 (intenseYear)	gehoa_2 (homeWorkYear)	getota_2 (totalYear)
M-HOO	↑W	$\overline{\uparrow \mathrm{W}}$	$\overline{\uparrow \mathrm{R}}$	↓M	↑M	$\overline{\uparrow \mathrm{R}}$	↑W	↑R	↓M	↑R
M-HB	↑W	↓R	↑R	$\overline{\downarrow \mathrm{W}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{W}}$	↓R	↑R	↓W	$\overline{\uparrow \mathrm{R}}$
M-BasicBak	$\overline{\uparrow \mathrm{R}}$	↓R	↑R	↓M	↑R	$\overline{\uparrow \mathrm{R}}$	↓R	↑R	↓R	$\overline{\uparrow \mathrm{W}}$
M-ProtoCaloric	$\overline{\uparrow \mathrm{R}}$	↑R	$\overline{\downarrow \mathrm{W}}$	↓R	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	↓M	↓R	$\overline{\downarrow \mathrm{R}}$
W-HOO	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↓W	$\overline{\uparrow \mathrm{R}}$	↓M	↓M	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↑W	↓W
W-HB	↑W	↓R	↓W	$\overline{\uparrow \mathrm{R}}$	↓R	$\overline{\uparrow \mathrm{W}}$	↓R	↓R	$\overline{\uparrow \mathrm{R}}$	↓R
W-BasicBak	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	↑R	↓W	↑R	$\overline{\downarrow \mathrm{R}}$	↑W	$\overline{\uparrow \mathrm{R}}$	↓W	$\overline{\uparrow \mathrm{R}}$
W-ProtoCaloric	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{W}}$	$\overline{\uparrow \mathrm{R}}$	↑M	$\overline{\uparrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↑M	↑M	$\overline{\uparrow \mathrm{R}}$
WM-HOO	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↓R	$\overline{\uparrow \mathrm{R}}$	↓R	$\overline{\downarrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↓R	$\overline{\uparrow \mathrm{R}}$	↓R
WM-HB	↑W	$\overline{\downarrow \mathrm{W}}$	↓R	$\overline{\uparrow \mathrm{R}}$	↓R	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↓R	$\overline{\uparrow \mathrm{R}}$	↓R
WM-BasicBak	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↓M	↑W	$\overline{\downarrow \mathrm{W}}$	$\overline{\uparrow \mathrm{M}}$	↓W	↓R	$\overline{\uparrow \mathrm{R}}$	↓M
WM-ProtoCaloric	↑W	$\overline{\uparrow \mathrm{R}}$	↓M	$\overline{\uparrow \mathrm{R}}$	$\overline{\downarrow \mathrm{R}}$	↑R	↑M	$\overline{\downarrow \mathrm{R}}$	$\overline{\uparrow \mathrm{R}}$	↑W

Table B.57: Descriptors for post partition P_f of Physical activity characteristics

B.6.2 Class Panel Graphs of partition P_f

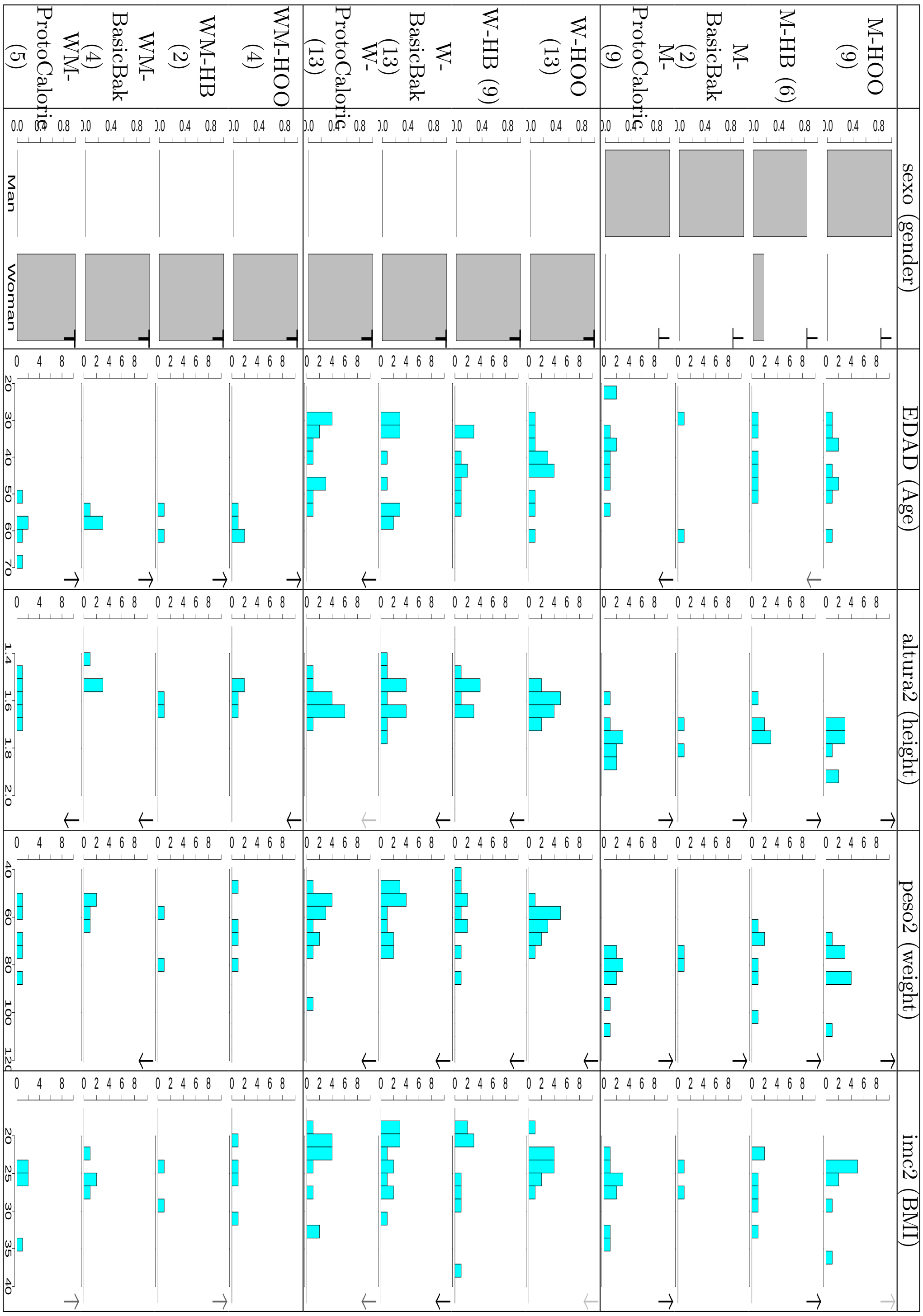


Table B.58: Biometrics characteristics $\sim$ Cross (Post) (a)

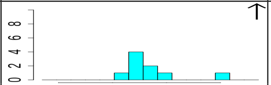
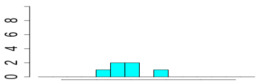
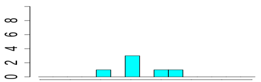
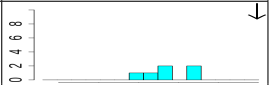
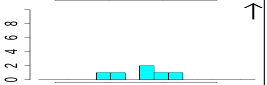
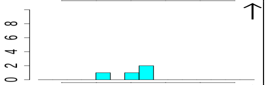
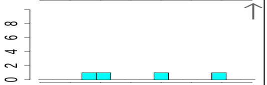
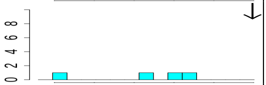
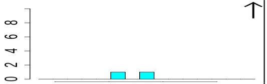
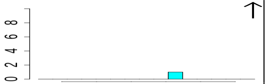
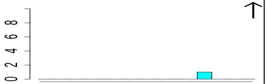
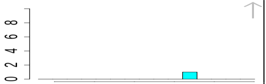
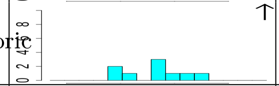
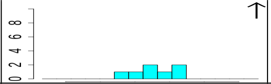
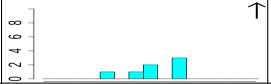
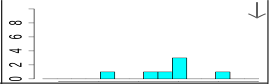
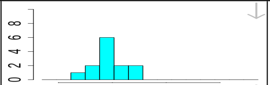
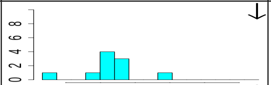
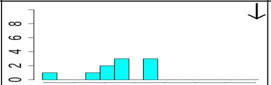
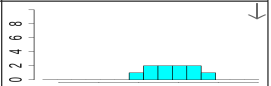
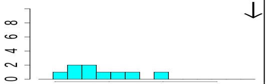
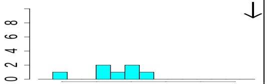
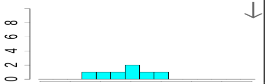
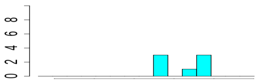
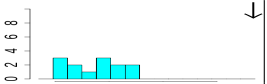
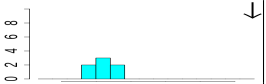
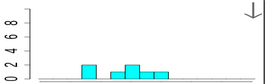
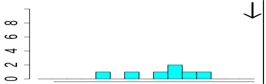
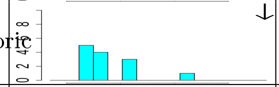
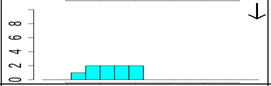
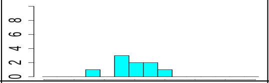
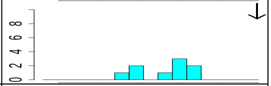
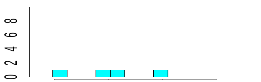
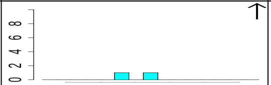
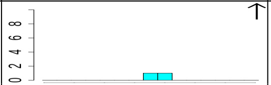
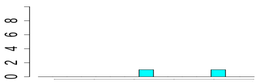
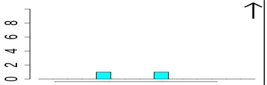
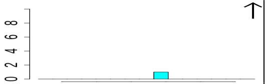
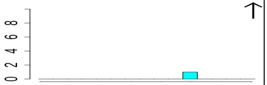
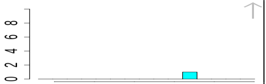
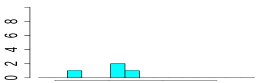
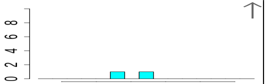
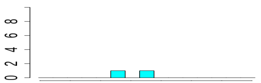
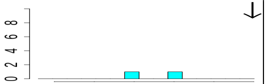
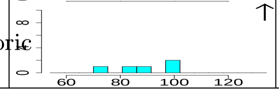
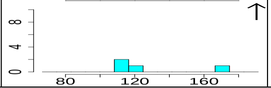
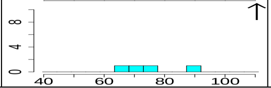
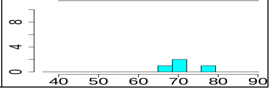
	cintura2 (waist:2)	pas_esds_fin (Systolic pressure)	pad_esds_fin (Diastolic pressure)	fc_fin (Heart Rate)
M-HOO (9)				
M-HB (6)				
M-BasicBak (2)				
M-ProtoCaloric (9)				
W-HOO (13)				
W-HB (9)				
W-BasicBak (13)				
W-ProtoCaloric (13)				
WM-HOO (4)				
WM-HB (2)				
WM-BasicBak (4)				
WM-ProtoCaloric (5)				

Table B.59: Biometrics characteristics ∼ Cross (Post) (b)

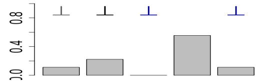
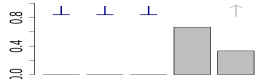
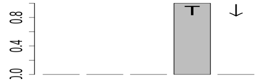
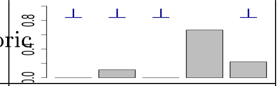
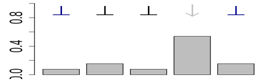
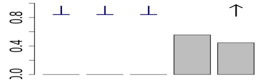
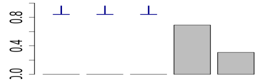
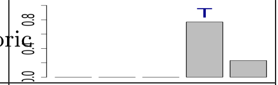
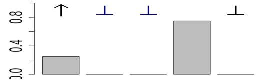
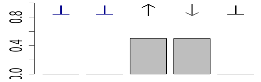
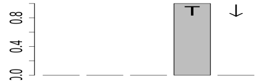
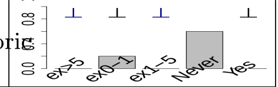
	tabaco2 (tobacco)
M-HOO (9)	
M-HB (6)	
M-BasicBak (2)	
M-ProtoCaloric (9)	
W-HOO (13)	
W-HB (9)	
W-BasicBak (13)	
W-ProtoCaloric (13)	
WM-HOO (4)	
WM-HB (2)	
WM-BasicBak (4)	
WM-ProtoCaloric (5)	

Table B.60: Tobacco characteristics ~ Cross (Post)

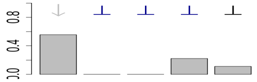
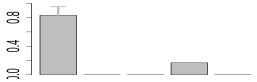
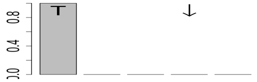
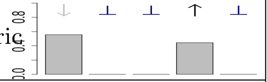
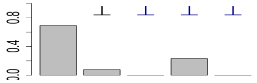
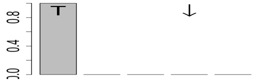
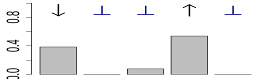
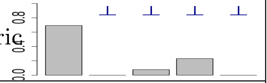
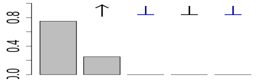
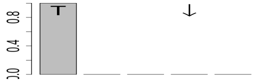
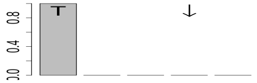
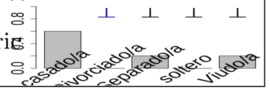
	est_civ2 (civilState)
M-HOO (9)	
M-HB (6)	
M-BasicBak (2)	
M-ProtoCaloric (9)	
W-HOO (13)	
W-HB (9)	
W-BasicBak (13)	
W-ProtoCaloric (13)	
WM-HOO (4)	
WM-HB (2)	
WM-BasicBak (4)	
WM-ProtoCaloric (5)	

Table B.61: Sociodemographic characteristics ∼ Cross (Post)

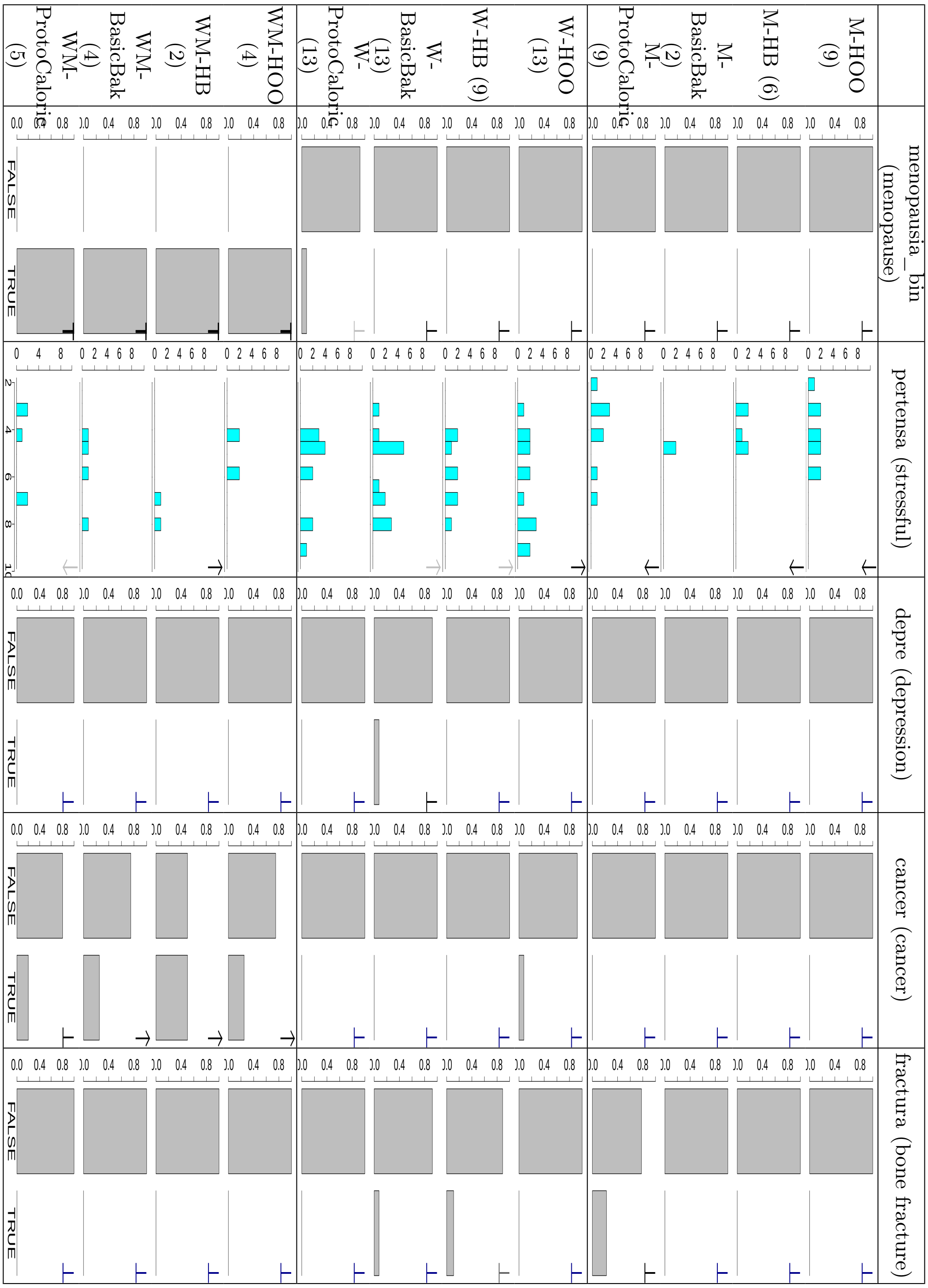


Table B.62: Menopause & Diseases characteristics ∼ Cross (Post) (a)

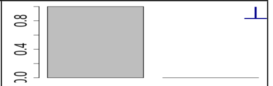
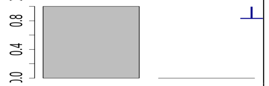
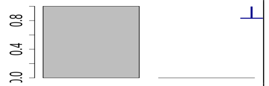
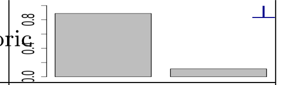
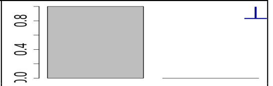
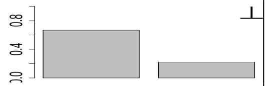
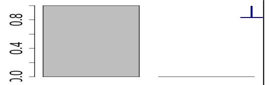
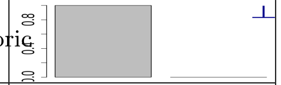
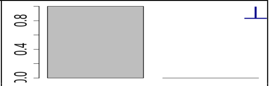
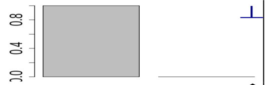
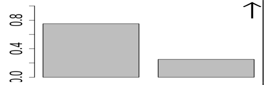
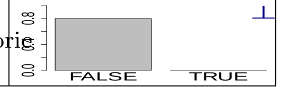
	disnea2 (dyspnea)
M-HOO (9)	
M-HB (6)	
M-BasicBak (2)	
M-ProtoCaloric (9)	
W-HOO (13)	
W-HB (9)	
W-BasicBak (13)	
W-ProtoCaloric (13)	
WM-HOO (4)	
WM-HB (2)	
WM-BasicBak (4)	
WM-ProtoCaloric (5)	

Table B.63: Menopause & Diseases characteristics ~ Cross (Post) (b)

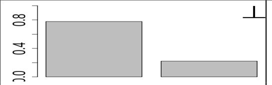
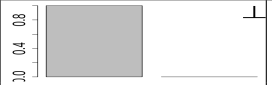
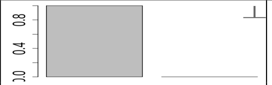
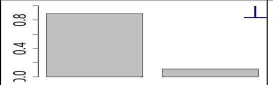
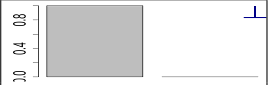
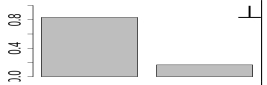
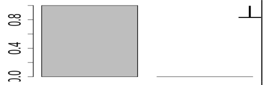
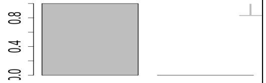
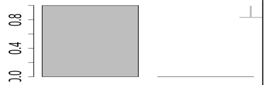
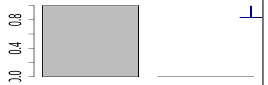
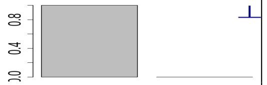
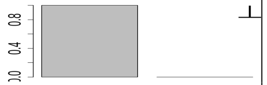
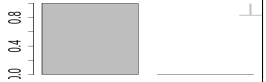
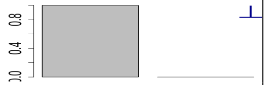
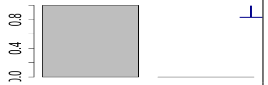
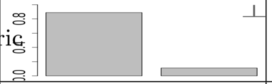
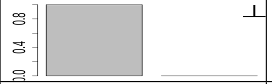
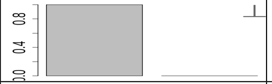
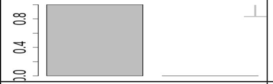
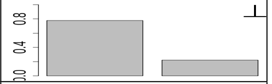
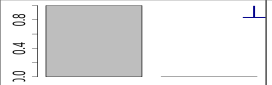
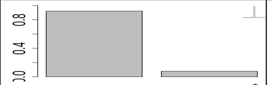
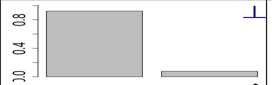
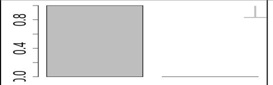
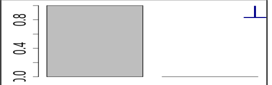
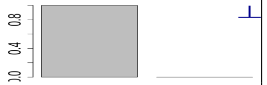
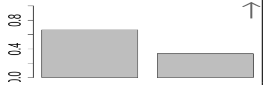
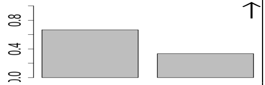
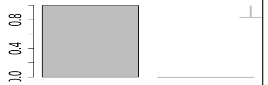
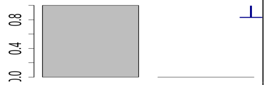
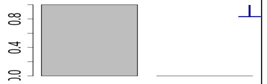
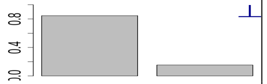
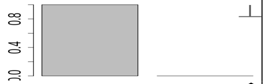
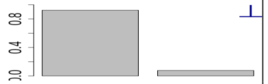
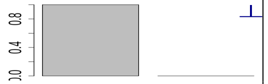
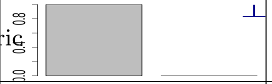
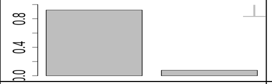
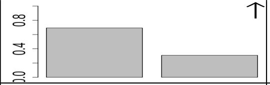
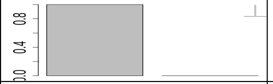
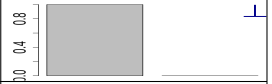
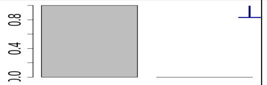
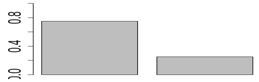
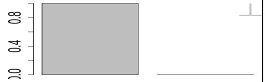
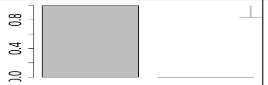
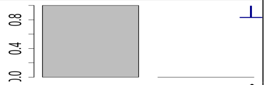
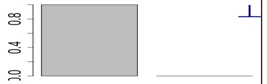
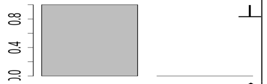
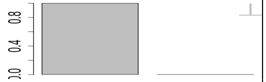
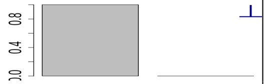
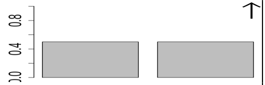
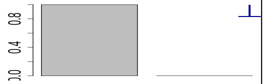
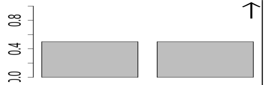
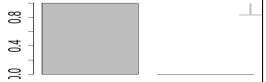
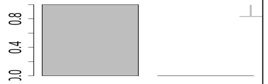
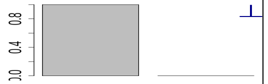
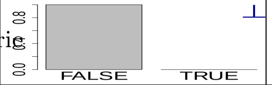
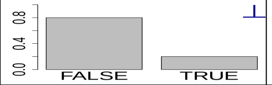
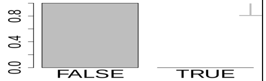
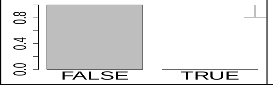
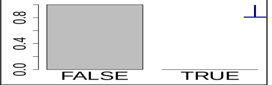
	maspirina2 (aspirin)	mdolor2 (NSAID)	mansiedad2 (anxiolytic)	mvitMin2 (vitamin or minerals)	mtension2 (heart)
M-HOO (9)					
M-HB (6)					
M-BasicBak (2)					
M-ProtoCaloric (9)					
W-HOO (13)					
W-HB (9)					
W-BasicBak (13)					
W-ProtoCaloric (13)					
WM-HOO (4)					
WM-HB (2)					
WM-BasicBak (4)					
WM-ProtoCaloric (5)					

Table B.64: Drugs characteristics ∼ Cross (Post) (a)

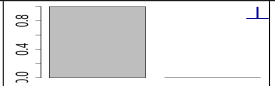
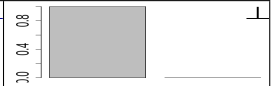
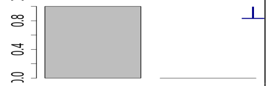
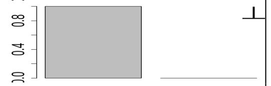
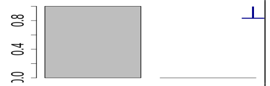
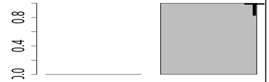
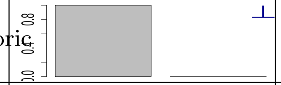
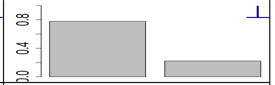
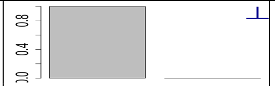
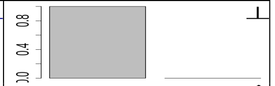
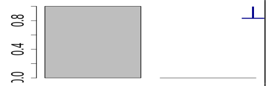
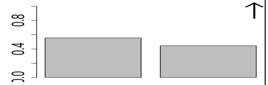
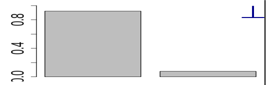
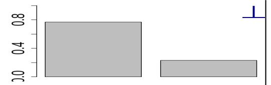
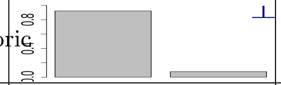
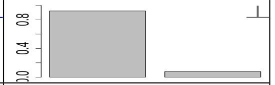
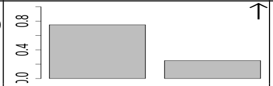
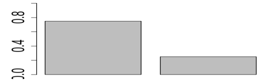
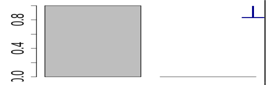
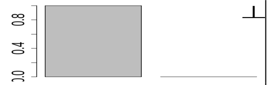
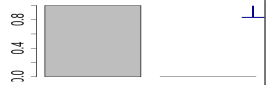
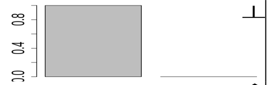
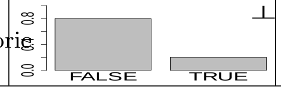
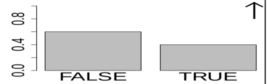
	mhormo2 (hormon)	motros2 (other)
M-HOO (9)		
M-HB (6)		
M-BasicBak (2)		
M-ProtoCaloric (9)		
W-HOO (13)		
W-HB (9)		
W-BasicBak (13)		
W-ProtoCaloric (13)		
WM-HOO (4)		
WM-HB (2)		
WM-BasicBak (4)		
WM-ProtoCaloric (5)		

Table B.65: Drugs characteristics ~ Cross (Post) (b)

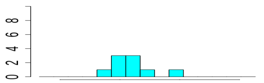
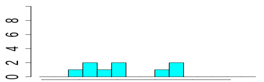
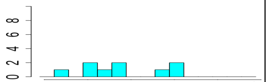
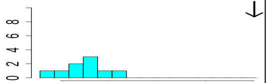
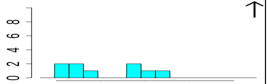
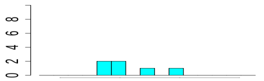
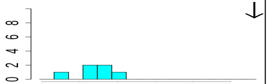
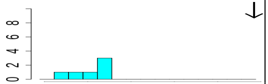
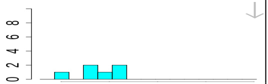
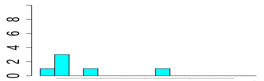
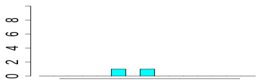
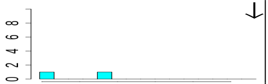
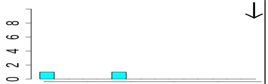
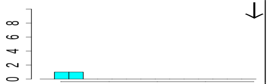
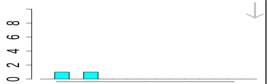
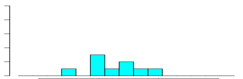
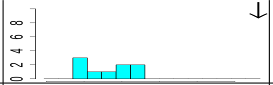
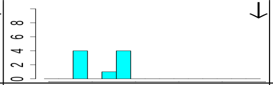
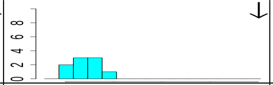
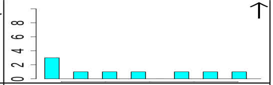
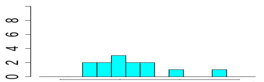
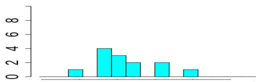
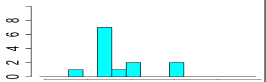
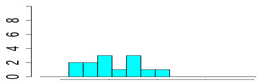
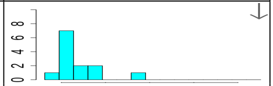
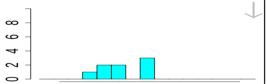
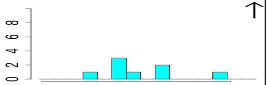
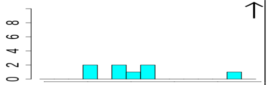
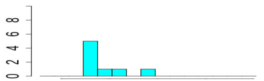
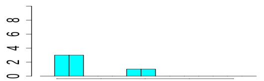
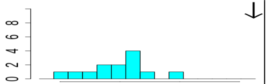
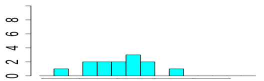
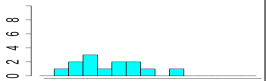
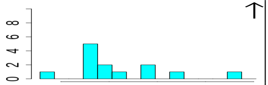
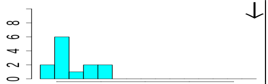
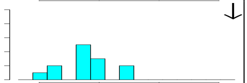
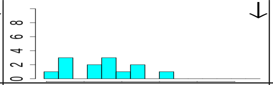
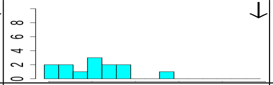
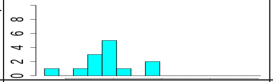
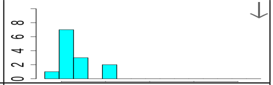
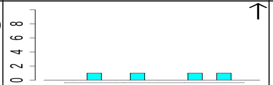
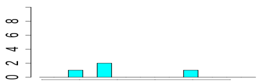
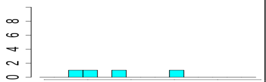
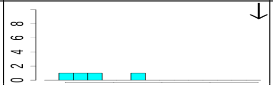
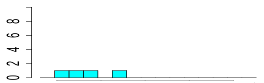
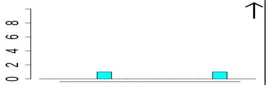
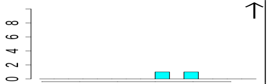
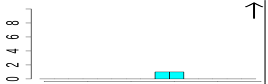
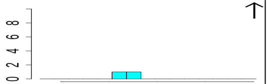
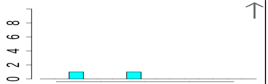
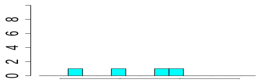
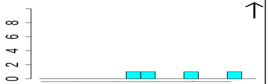
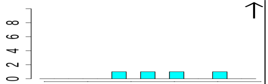
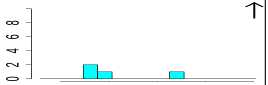
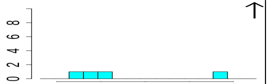
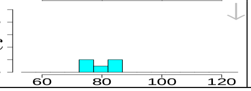
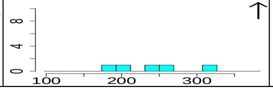
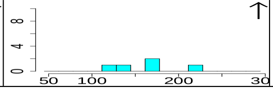
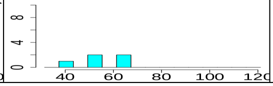
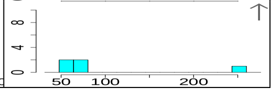
	gluc3 (glucose)	cholest1 (cholesterol)	LDL_3 (LDL)	hdl3 (HDL)	tryg3 (triglycerides)
M-HOO (9)					
M-HB (6)					
M-BasicBak (2)					
M-ProtoCaloric (9)					
W-HOO (13)					
W-HB (9)					
W-BasicBak (13)					
W-ProtoCaloric (13)					
WM-HOO (4)					
WM-HB (2)					
WM-BasicBak (4)					
WM-ProtoCaloric (5)					

Table B.66: Biomarkers characteristics ~ Cross (Post) (a)

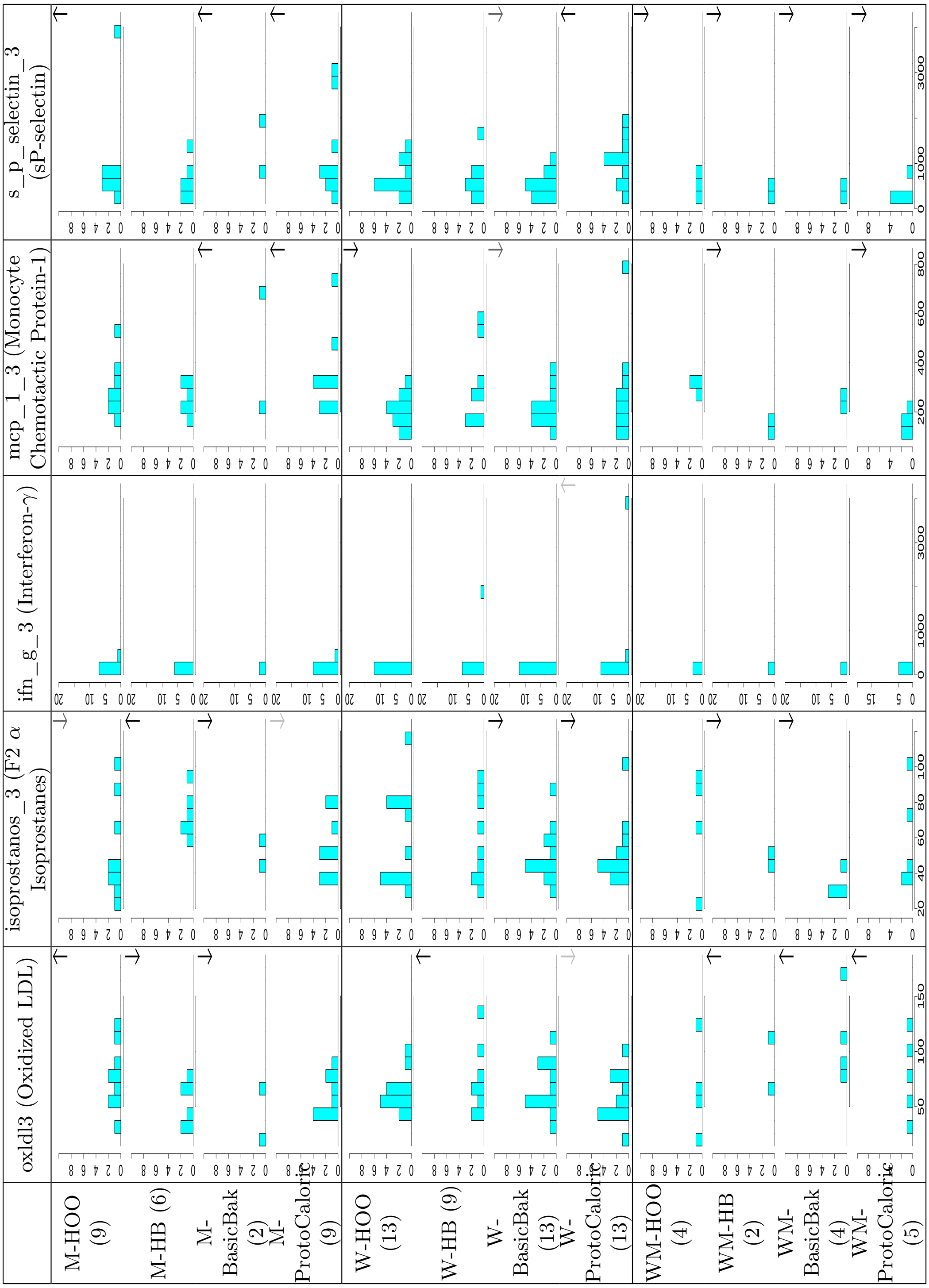


Table B.67: Biomarkers characteristics ∼ Cross (Post) (b)

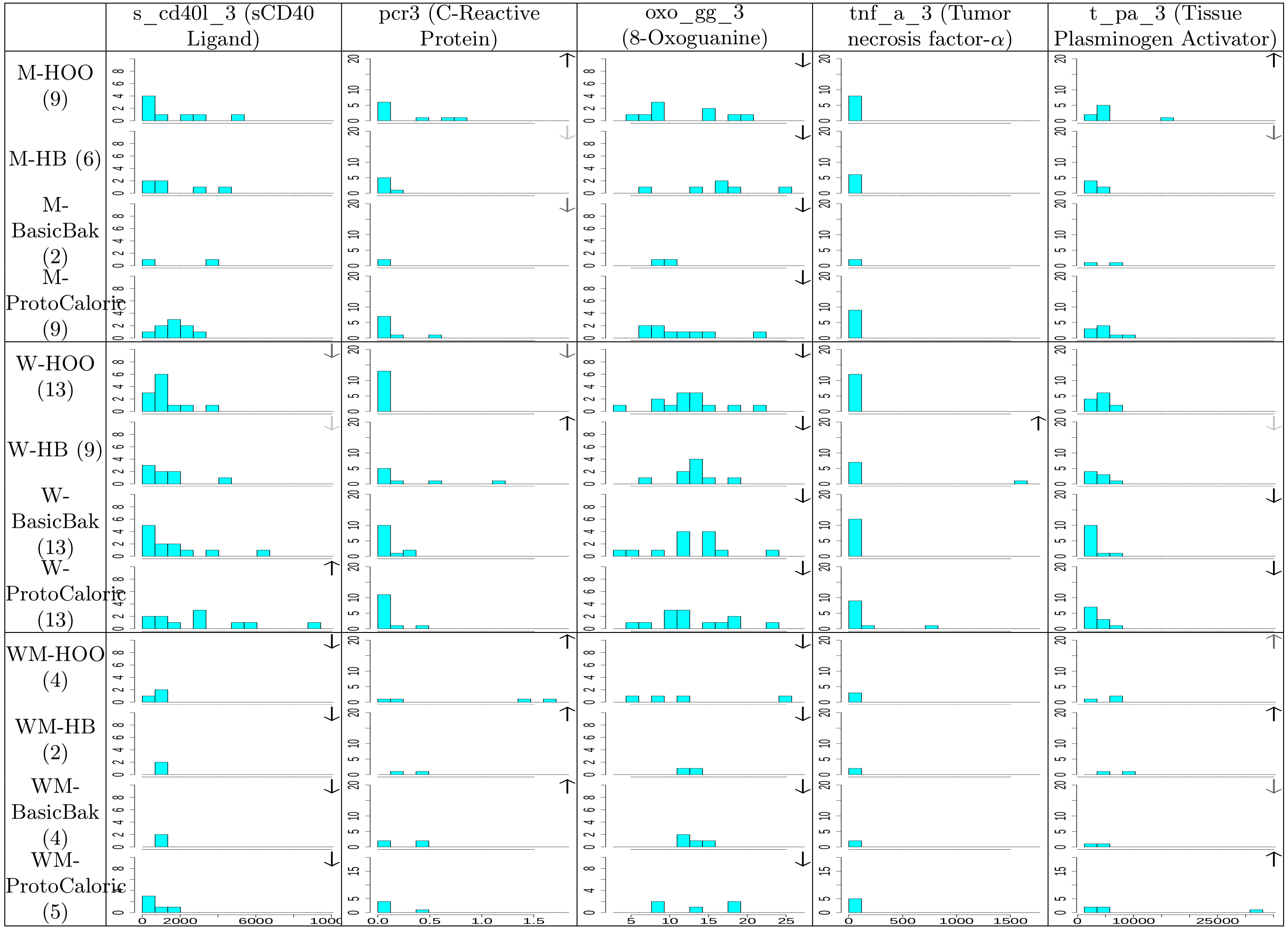


Table B.68: Biomarkers characteristics ∼ Cross (Post) (c)

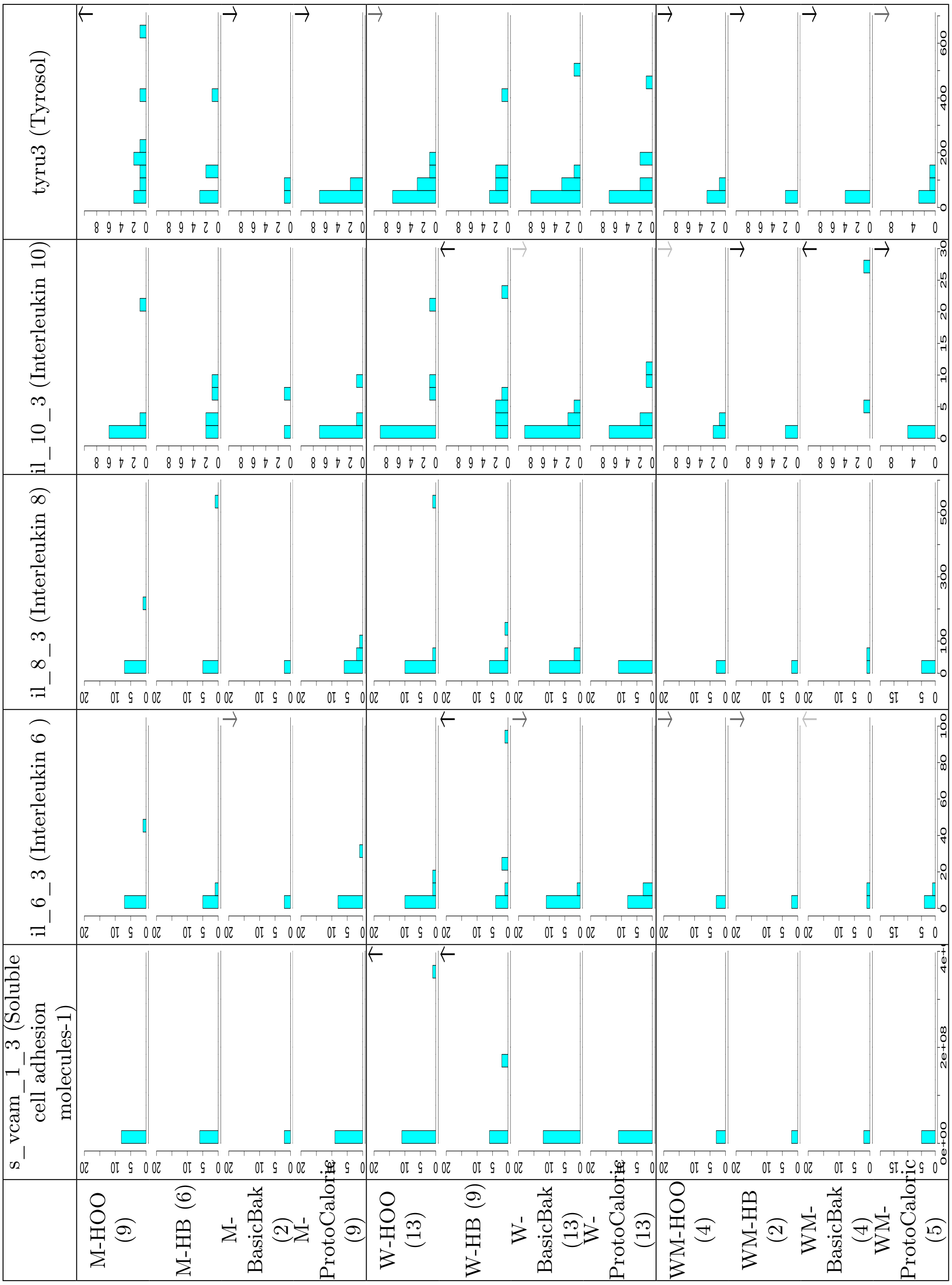


Table B.69: Biomarkers characteristics ~ Cross (Post) (d)

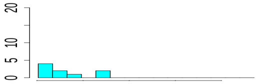
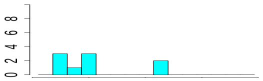
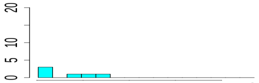
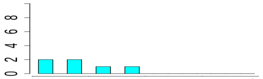
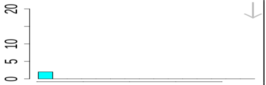
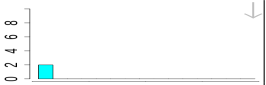
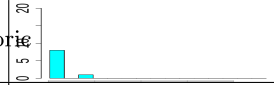
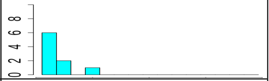
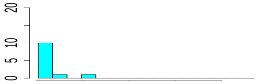
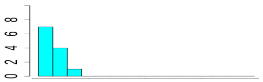
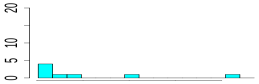
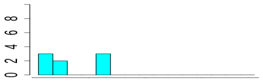
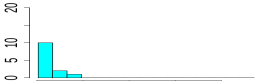
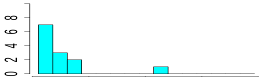
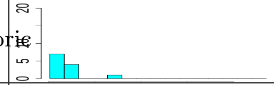
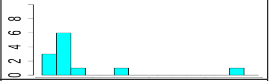
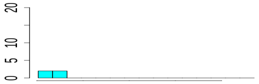
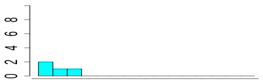
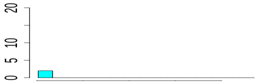
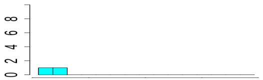
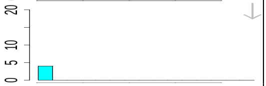
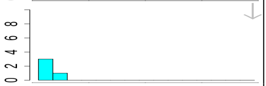
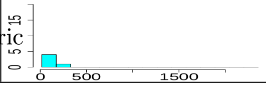
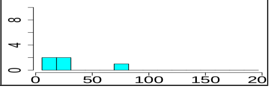
	ohtyru3 (OHTyrosol)	mohtyu3 (MOH_Tyrosol)
M-HOO (9)		
M-HB (6)		
M-BasicBak (2)		
M-ProtoCaloric (9)		
W-HOO (13)		
W-HB (9)		
W-BasicBak (13)		
W-ProtoCaloric (13)		
WM-HOO (4)		
WM-HB (2)		
WM-BasicBak (4)		
WM-ProtoCaloric (5)		

Table B.70: Biomarkers characteristics ∼ Cross (Post) (e)

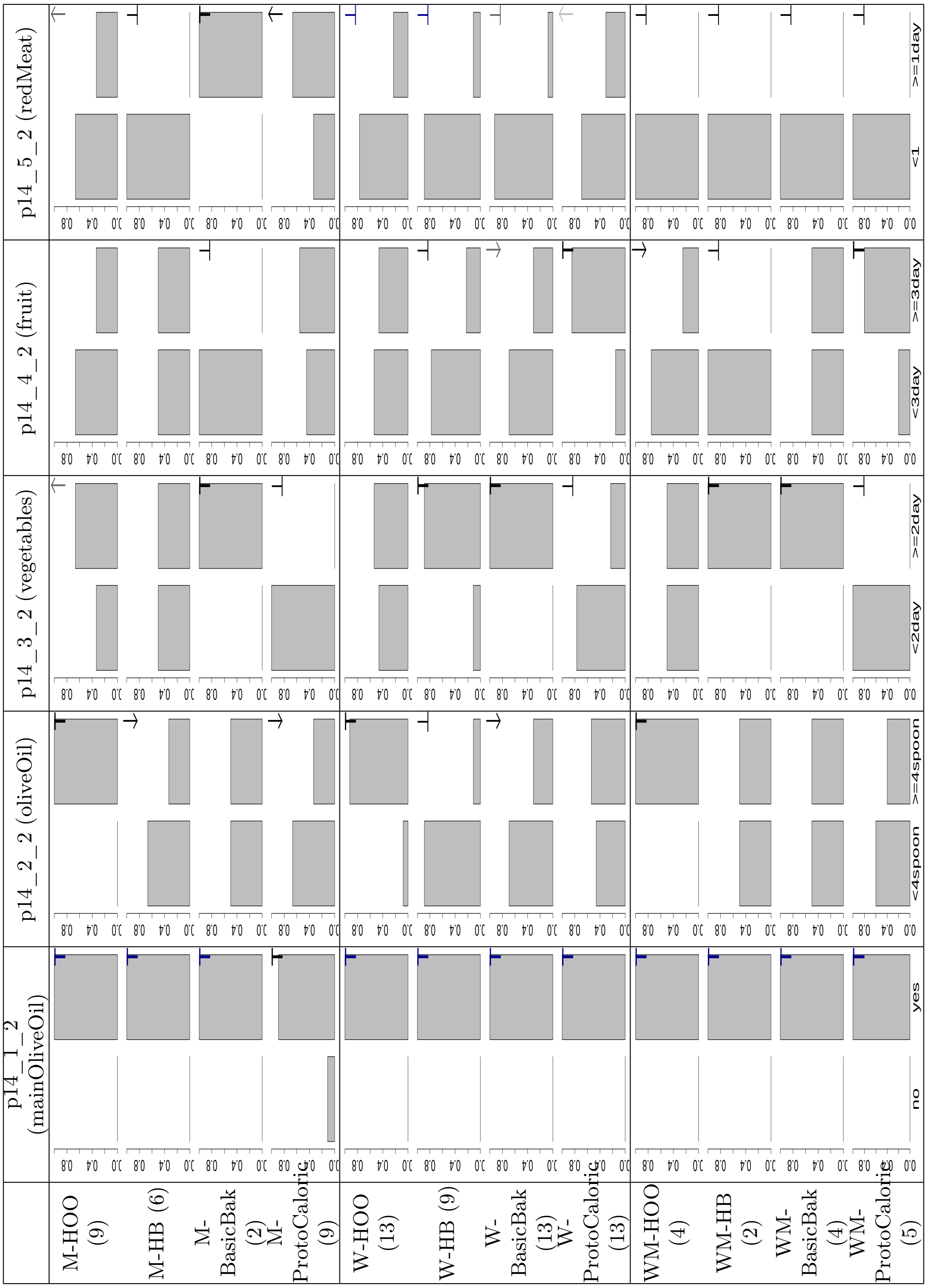


Table B.71: Diet characteristics ~ Cross (Pre) (a)

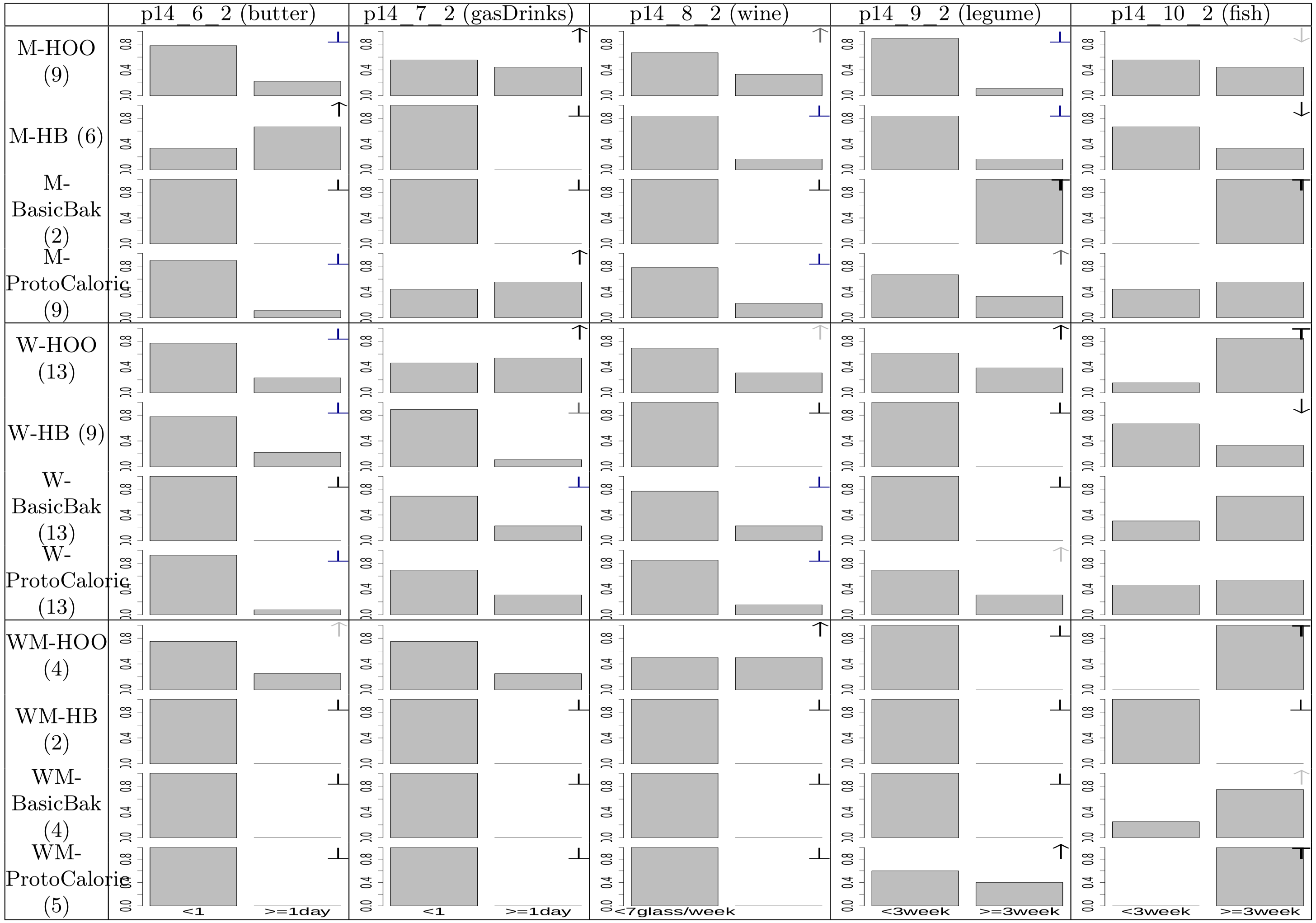


Table B.72: Diet characteristics ~ Cross (Pre) (b)

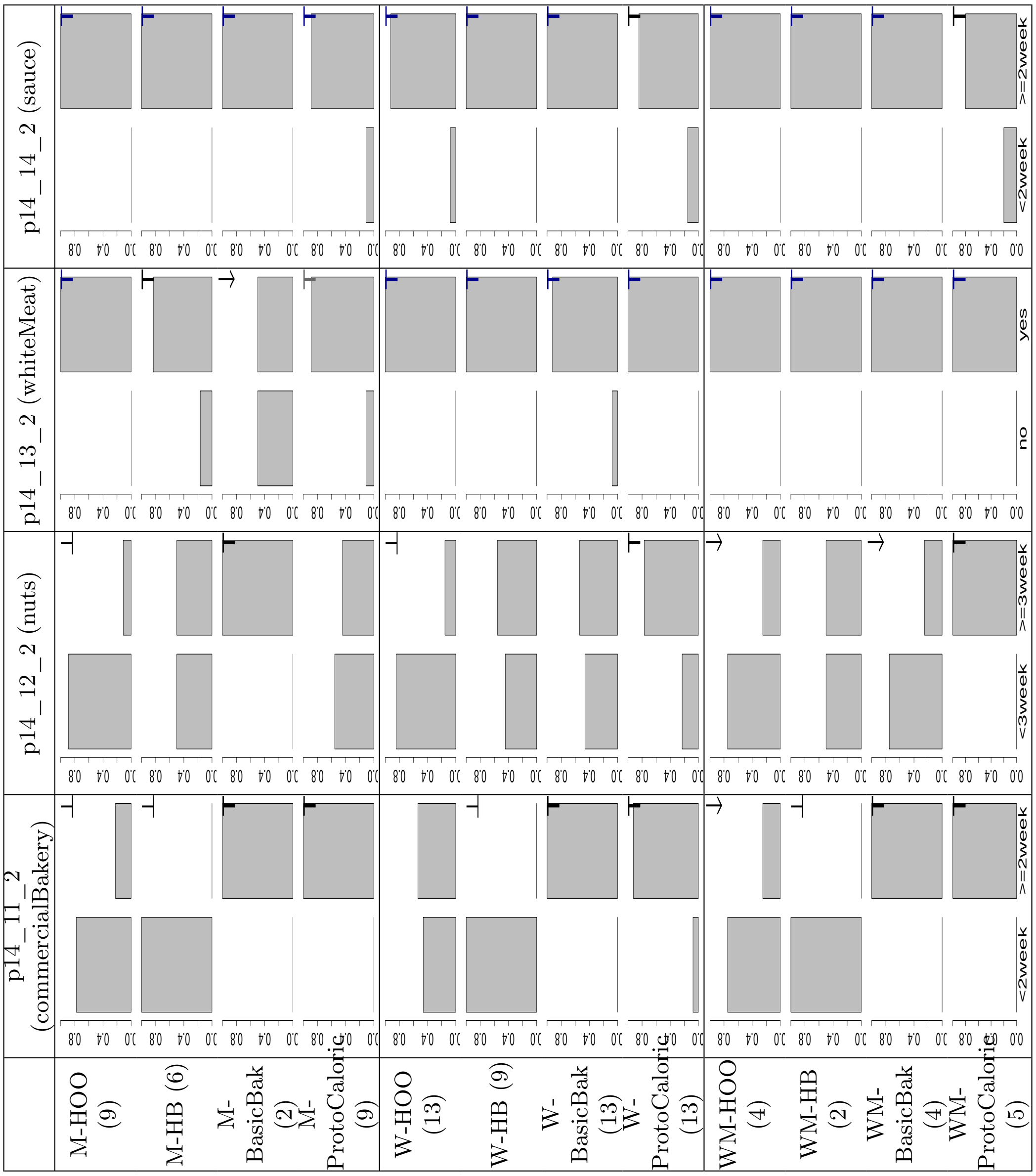


Table B.73: Diet characteristics ∼ Cross (Pre) (c)

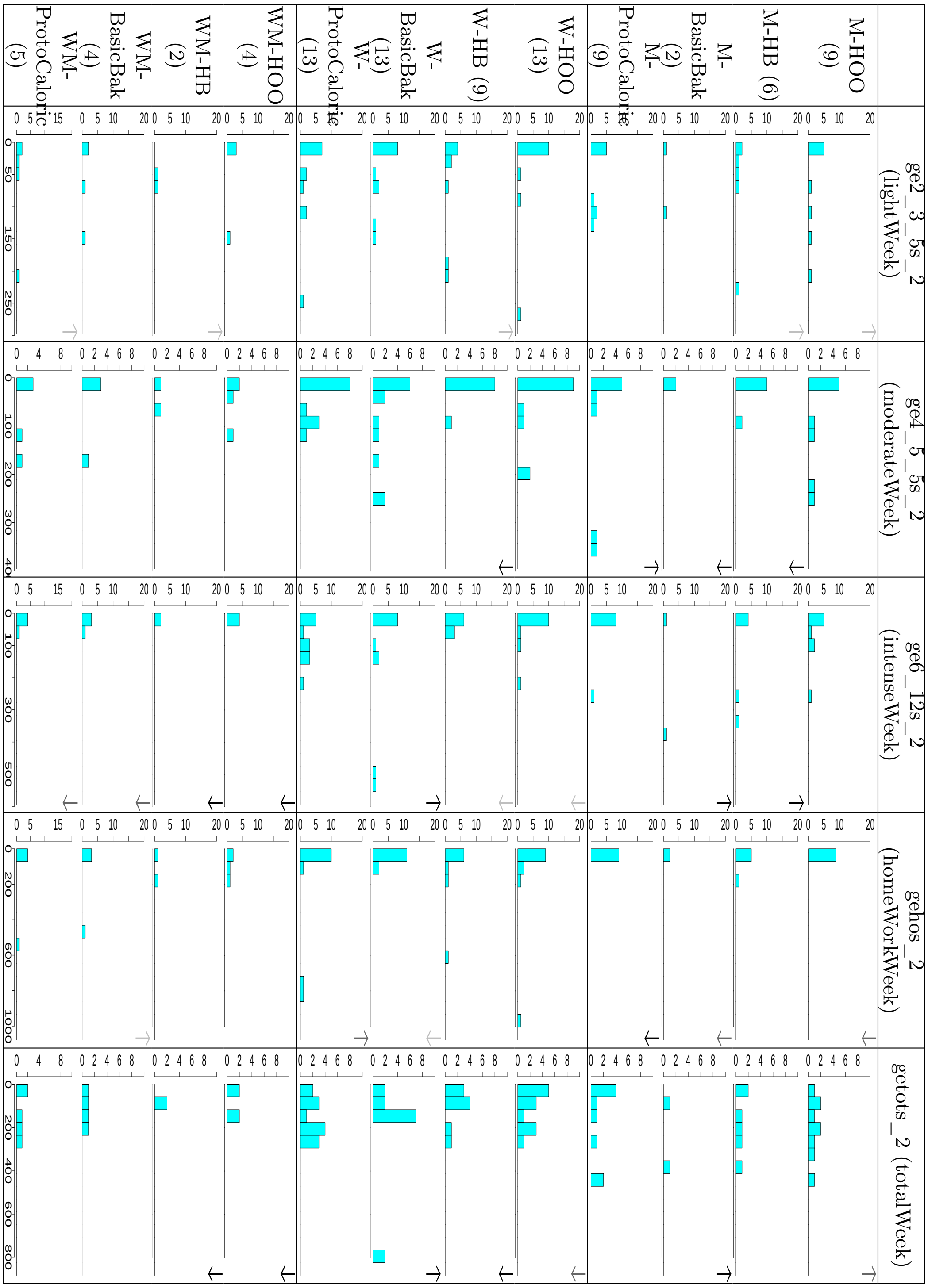


Table B.74: Physical activity characteristics ∼ Cross (Pre) (a)

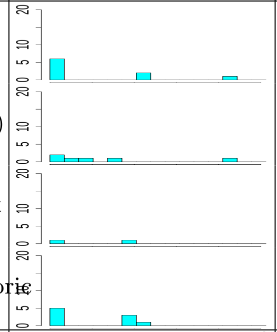
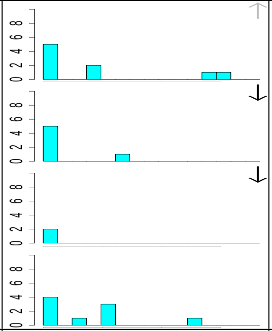
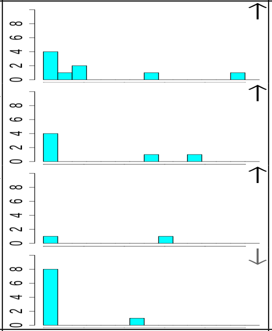
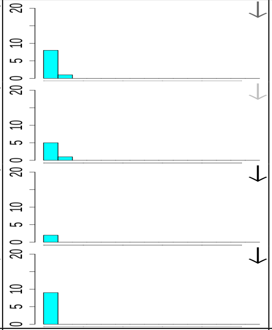
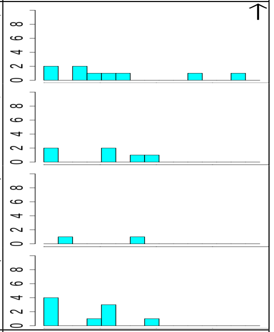
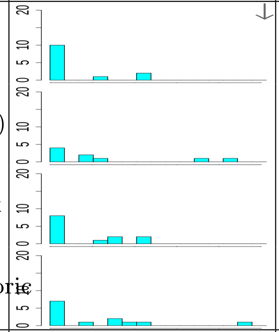
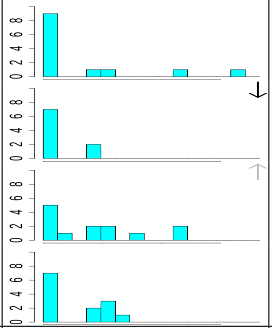
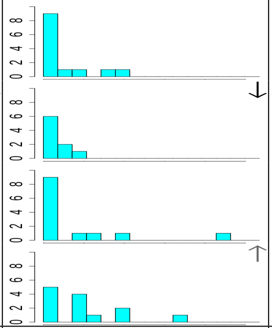
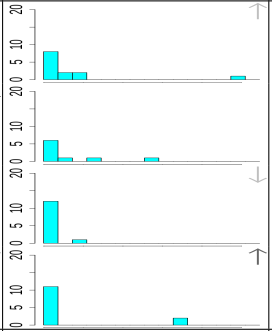
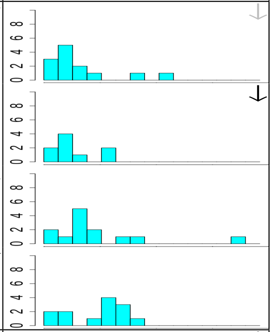
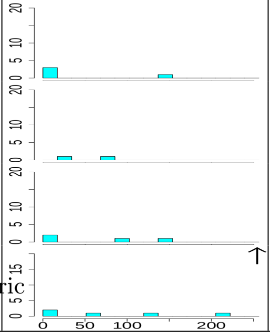
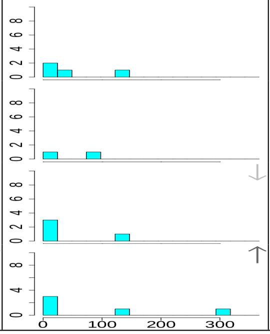
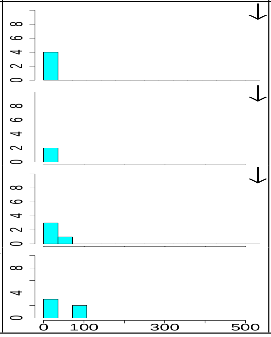
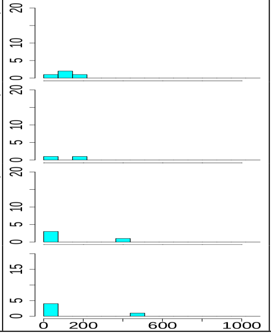
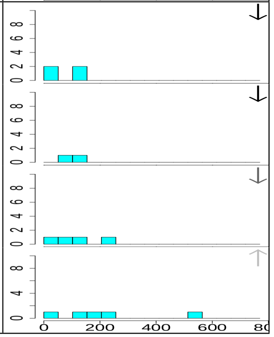
	ge2_3_5a_2 (lightYear)	ge4_5_5a_2 (moderateYear)	ge6_12a_2 (intenseYear)	gehoa_2 (homeWorkYear)	getota_2 (totalYear)
M-HOO (9) M-HB (6) M-BasicBak (2) M-ProtoCaloric (9)					
W-HOO (13) W-HB (9) W-BasicBak (13) W-ProtoCaloric (13)					
WM-HOO (4) WM-HB (2) WM-BasicBak (4) WM-ProtoCaloric (5)					

Table B.75: Physical activity characteristics ∼ Cross (Pre) (b)

B.6.3 Automatic Profiles of P_f

M-HOO : **Less** Woman (-H) in *sexo* (gender); *altura2* (height) is higher(H); *peso2* (weight) is higher(H); *imc2* (BMI) is higher(HW); *cintura2* (waist:2) is higher(H); *pad_esds_fin* (Diastolic pressure) is (WI); *fc_fin* (Heart Rate) is lower(-H); More ex>5 (SM) and more ex0-1 (HW) in *tabaco2* (tobacco); Less casado/a () and more Viudo/a (S) in *est_civ2* (civilState); **Less** TRUE (-H) in *menopausia_bin* (menopause); *pertensa* (stressful) is lower(-H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); **Less** TRUE (-H) in *mdolor2* (NSAID); **Less** TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); **Less** TRUE (-H) in *motros2* (other); *hdl3* (HDL) is lower(-H); *tryg3* (triglycerides) is higher(H); *oxldl3* (Oxidized LDL) is higher(S); *isoprostanos_3* (F2 α Isoprostanes) is lower(-SM); *s_p_selectin_3* (sP-selectin) is higher(H); *pcr3* (C-Reactive Protein) is higher(S); *oxo_gg_3* (8-Oxoguanine) is lower(-H); *t_pa_3* (Tissue Plasminogen Activator) is higher(S); *tyru3* (Tyrosol) is higher(S); More yes (H) in *p14_1_2* (mainOliveOil); **More** >=4spoon (H) in *p14_2_2* (oliveOil); More >=2day (SM) in *p14_3_2* (vegetables); More >=1day (HM) in *p14_5_2* (redMeat); Less >=1day (-H) in *p14_6_2* (butter); More >=1day (H) in *p14_7_2* (gasDrinks); More >=7glass/week (HM) in *p14_8_2* (wine); Less >=3week (-H) in *p14_9_2* (legume); Less >=3week (-HL) in *p14_10_2* (fish); **Less** >=2week (-H) in *p14_11_2* (commercialBakery); **Less** >=3week (-H) in *p14_12_2* (nuts); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce); *ge2_3_5s_2* (lightWeek) is higher(); *gehos_2* (homeWorkWeek) is lower(-SM); *getots_2* (totalWeek) is higher(SM); *ge4_5_5a_2* (moderateYear) is higher(); *ge6_12a_2* (intenseYear) is higher(S); *gehoa_2* (homeWorkYear) is lower(-SM); *getota_2* (totalYear) is higher(S);

M-HB : **Less** Woman (-H) in *sexo* (gender); *EDAD* (Age) is lower(-HW); *altura2* (height) is higher(H); *peso2* (weight) is higher(H); *imc2* (BMI) is higher(H); *cintura2* (waist:2) is higher(H); *pas_esds_fin* (Systolic pressure) is higher(H); *pad_esds_fin* (Diastolic pressure) is higher(HM); *fc_fin* (Heart Rate) is lower(-H); More Yes () in *tabaco2* (tobacco); More casado/a () in *est_civ2* (civilState); **Less** TRUE (-H) in *menopausia_bin* (menopause); *pertensa* (stressful) is lower(-H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); **Less** TRUE (-H) in *mdolor2* (NSAID); **Less** TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); **Less** TRUE (-H) in *motros2* (other); *cholest1* (cholesterol) is lower(-H); *LDL_3* (LDL) is lower(-H); *hdl3* (HDL) is lower(-HW); *oxldl3* (Oxidized LDL) is lower(-S); *isoprostanos_3* (F2 α Isoprostanes) is higher(S); *s_p_selectin_3* (sP-selectin) is (-WI); *pcr3* (C-Reactive Protein) is lower(); *oxo_gg_3* (8-Oxoguanine) is lower(-H); *t_pa_3* (Tissue Plasminogen Activator) is lower(-SM); More yes (H) in *p14_1_2* (mainOliveOil); Less >=4spoon (-H) in *p14_2_2* (oliveOil); **Less** >=1day (-H) in *p14_5_2* (redMeat); More >=1day (H) in *p14_6_2* (butter); **Less** >=1day (-H) in *p14_7_2* (gasDrinks); Less >=7glass/week (-H) in *p14_8_2* (wine); Less >=3week (-H) in *p14_9_2* (legume); Less >=3week (-H) in *p14_10_2* (fish); **Less** >=2week (-H) in *p14_11_2* (commercialBakery); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce); *ge2_3_5s_2* (lightWeek) is higher(HL); *ge4_5_5s_2* (moderateWeek) is lower(-H); *ge6_12s_2* (intenseWeek) is higher(S); *ge4_5_5a_2* (moderateYear) is lower(-H); *ge6_12a_2* (intenseYear) is higher(S); *gehoa_2* (homeWorkYear) is lower();

M-BasicBak : **Less** Woman (-H) in *sexo* (gender); *altura2* (height) is higher(H); *peso2* (weight) is higher(H); *cintura2* (waist:2) is higher(H); *pas_esds_fin* (Systolic pressure) is higher(H); *pad_esds_fin* (Diastolic pressure) is higher(H); *fc_fin* (Heart Rate) is higher(HW); More Never (S) and less Yes (-S) in *tabaco2* (tobacco); More casado/a (S) and less soltero (-S) in *est_civ2* (civilState); **Less** TRUE (-H) in *menopausia_bin* (menopause); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); **Less** TRUE (-H) in *mdolor2* (NSAID); Less TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); **More**

TRUE (H) in *motros2* (other); *cholest1* (cholesterol) is lower(-H); *LDL_3* (LDL) is lower(-H); *hdl3* (HDL) is lower(-H); *tryg3* (triglycerides) is lower(-HW); *oxldl3* (Oxidized LDL) is lower(-S); *isoprostanos_3* (F2 α Isoprostanes) is lower(-S); *mcp_1_3* (Monocyte Chemotactic Protein-1) is higher(SM); *s_p_selectin_3* (sP-selectin) is higher(H); *pcr3* (C-Reactive Protein) is lower(-SM); *oxo_gg_3* (8-Oxoguanine) is lower(-H); *il_6_3* (Interleukin 6) is lower(-SM); *tyru3* (Tyrosol) is lower(-S); *ohtyru3* (OHTyrosol) is lower(); *mohtyu3* (MOH_Tyrosol) is lower(); More yes (H) in *p14_1_2* (mainOliveOil); **More** >=2day (H) in *p14_3_2* (vegetables); **Less** >=3day (-HM) in *p14_4_2* (fruit); **More** >=1day (H) in *p14_5_2* (redMeat); **Less** >=1day (-H) in *p14_6_2* (butter); **Less** >=1day (-H) in *p14_7_2* (gasDrinks); **Less** >=7glass/week (-H) in *p14_8_2* (wine); **More** >=3week (H) in *p14_9_2* (legume); **More** >=3week (HW) in *p14_10_2* (fish); **More** >=2week (H) in *p14_11_2* (commercialBakery); **More** >=3week (S) in *p14_12_2* (nuts); Less yes (-H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce); *ge4_5_5s_2* (moderateWeek) is lower(-S); *ge6_12s_2* (intenseWeek) is higher(H); *gehos_2* (homeWorkWeek) is lower(-SM); *getots_2* (totalWeek) is higher(HM); *ge4_5_5a_2* (moderateYear) is lower(-S); *ge6_12a_2* (intenseYear) is higher(S); *gehoa_2* (homeWorkYear) is lower(-S);

M-ProtoCaloric : **Less** Woman (-H) in *sexo* (gender); *EDAD* (Age) is lower(-HM); *altura2* (height) is higher(H); *peso2* (weight) is higher(H); *imc2* (BMI) is higher(H); *cintura2* (waist:2) is higher(H); *pas_esds_fin* (Systolic pressure) is higher(H); *pad_esds_fin* (Diastolic pressure) is higher(H); *fc_fin* (Heart Rate) is lower(-HM); Less casado/a () and more soltero (S) in *est_civ2* (civilState); **Less** TRUE (-H) in *menopausia_bin* (menopause); *pertensa* (stressful) is lower(-H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); **Less** TRUE (-H) in *mdolor2* (NSAID); **Less** TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); Less TRUE (-H) in *motros2* (other); *cholest1* (cholesterol) is lower(-H); *LDL_3* (LDL) is lower(-H); *hdl3* (HDL) is lower(-H); *tryg3* (triglycerides) is higher(H); *isoprostanos_3* (F2 α Isoprostanes) is lower(); *mcp_1_3* (Monocyte Chemotactic Protein-1) is higher(SM); *s_p_selectin_3* (sP-selectin) is higher(H); *oxo_gg_3* (8-Oxoguanine) is lower(-H); *tyru3* (Tyrosol) is lower(-S); More yes (H) in *p14_1_2* (mainOliveOil); Less >=4spoon (-S) in *p14_2_2* (oliveOil); **Less** >=2day (-H) in *p14_3_2* (vegetables); More >=1day (H) in *p14_5_2* (redMeat); Less >=1day (-H) in *p14_6_2* (butter); More >=1day (S) in *p14_7_2* (gasDrinks); Less >=7glass/week (-H) in *p14_8_2* (wine); More >=3week (HM) in *p14_9_2* (legume); **More** >=2week (H) in *p14_11_2* (commercialBakery); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce); *ge4_5_5s_2* (moderateWeek) is higher(S); *gehos_2* (homeWorkWeek) is lower(-S); *ge6_12a_2* (intenseYear) is lower(-SM); *gehoa_2* (homeWorkYear) is lower(-S);

W-HOO : **More** Woman (H) in *sexo* (gender); *altura2* (height) is (-WI); *peso2* (weight) is lower(-H); *imc2* (BMI) is lower(-HW); *cintura2* (waist:2) is lower(-HW); *pas_esds_fin* (Systolic pressure) is lower(-H); *pad_esds_fin* (Diastolic pressure) is lower(-H); *fc_fin* (Heart Rate) is lower(-HM); More ex0-1 (S) and more ex1-5 (S) and less Never () in *tabaco2* (tobacco); More Divorciado/a (S) in *est_civ2* (civilState); **Less** TRUE (-H) in *menopausia_bin* (menopause); *pertensa* (stressful) is higher(H); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); Less TRUE (-H) in *mdolor2* (NSAID); Less TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); **Less** TRUE (-H) in *motros2* (other); *tryg3* (triglycerides) is lower(-HM); *mcp_1_3* (Monocyte Chemotactic Protein-1) is lower(-SW); *s_cd40l_3* (sCD40 Ligand) is lower(-SM); *pcr3* (C-Reactive Protein) is lower(-SM); *oxo_gg_3* (8-Oxoguanine) is lower(-H); *s_vcam_1_3* (Soluble cell adhesion molecules-1) is higher(HW); *tyru3* (Tyrosol) is lower(-SW); More yes (H) in *p14_1_2* (mainOliveOil); **More** >=4spoon (H) in *p14_2_2* (oliveOil); Less >=1day (-H) in *p14_5_2* (redMeat); Less >=1day (-H) in *p14_6_2* (butter); More >=1day (H) in *p14_7_2* (gasDrinks); More >=7glass/week (HW) in *p14_8_2* (wine); More >=3week (H) in *p14_9_2* (legume); **More** >=3week (HM) in *p14_10_2* (fish); >=2week (-WI) in *p14_11_2* (commercialBakery); **Less** >=3week (-H) in *p14_12_2* (nuts); More yes (H) in *p14_13_2* (white-

Meat); More >=2week (H) in *p14_14_2* (sauce); *ge6_12s_2* (intenseWeek) is lower(); *getots_2* (totalWeek) is lower(-SM); *ge2_3_5a_2* (lightYear) is lower(-SM); *gehoa_2* (homeWorkYear) is higher(); *getota_2* (totalYear) is lower();

W-HB : **More** Woman (H) in *sexo* (gender); *altura2* (height) is lower(-H); *peso2* (weight) is lower(-H); *cintura2* (waist:2) is lower(-H); *pas_esds_fin* (Systolic pressure) is lower(-H); *pad_esds_fin* (Diastolic pressure) is lower(-HM); More Yes (S) in *tabaco2* (tobacco); More casado/a (S) and less soltero (-S) in *est_civ2* (civilState); **Less** TRUE (-H) in *menopausia_bin* (menopause); *pertensa* (stressful) is higher(HW); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); More TRUE (HM) in *mdolor2* (NSAID); More TRUE (H) in *mansiedad2* (anxiolytic); **Less** TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); More TRUE (H) in *motros2* (other); *gluc3* (glucose) is lower(-HW); *cholest1* (cholesterol) is higher(S); *LDL_3* (LDL) is higher(S); *oxldl3* (Oxidized LDL) is higher(S); *s_cd40l_3* (sCD40 Ligand) is lower(); *pcr3* (C-Reactive Protein) is higher(S); *oxo_gg_3* (8-Oxoguanine) is lower(-H); *tnf_a_3* (Tumor necrosis factor-α) is higher(S); *t_pa_3* (Tissue Plasminogen Activator) is lower(-HW); *s_vcam_1_3* (Soluble cell adhesion molecules-1) is higher(HW); *il_6_3* (Interleukin 6) is higher(S); *il_10_3* (Interleukin 10) is higher(S); More yes (H) in *p14_1_2* (mainOliveOil); **Less** >=4spoon (-H) in *p14_2_2* (oliveOil); **More** >=2day (H) in *p14_3_2* (vegetables); **Less** >=3day (-HM) in *p14_4_2* (fruit); Less >=1day (-H) in *p14_5_2* (redMeat); Less >=1day (-H) in *p14_6_2* (butter); **Less** >=1day (-H) in *p14_7_2* (gasDrinks); **Less** >=7glass/week (-H) in *p14_8_2* (wine); **Less** >=3week (-H) in *p14_9_2* (legume); Less >=3week (-H) in *p14_10_2* (fish); **Less** >=2week (-H) in *p14_11_2* (commercialBakery); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce); *ge2_3_5s_2* (lightWeek) is higher(HL); *ge4_5_5s_2* (moderateWeek) is lower(-H); *ge6_12s_2* (intenseWeek) is lower(); *getots_2* (totalWeek) is lower(-S); *ge4_5_5a_2* (moderateYear) is lower(-H); *ge6_12a_2* (intenseYear) is lower(-S); *getota_2* (totalYear) is lower(-SM);

W-BasicBak : **More** Woman (H) in *sexo* (gender); *altura2* (height) is lower(-H); *peso2* (weight) is lower(-H); *imc2* (BMI) is lower(-H); *cintura2* (waist:2) is lower(-H); *pas_esds_fin* (Systolic pressure) is lower(-H); *pad_esds_fin* (Diastolic pressure) is lower(-HM); *fc_fin* (Heart Rate) is lower(-H); Less casado/a (-S) and more soltero (S) in *est_civ2* (civilState); **Less** TRUE (-H) in *menopausia_bin* (menopause); *pertensa* (stressful) is higher(HW); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); Less TRUE (-H) in *mdolor2* (NSAID); **Less** TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); Less TRUE (-H) in *motros2* (other); *gluc3* (glucose) is lower(-H); *hdl3* (HDL) is higher(S); *tryg3* (triglycerides) is lower(-H); *isoprostanos_3* (F2 α Isoprostanes) is lower(-H); *mcp_1_3* (Monocyte Chemotactic Protein-1) is lower(-SL); *s_p_selectin_3* (sP-selectin) is lower(-SM); *oxo_gg_3* (8-Oxoguanine) is lower(-H); *t_pa_3* (Tissue Plasminogen Activator) is lower(-H); *il_6_3* (Interleukin 6) is lower(-SM); *il_10_3* (Interleukin 10) is lower(); More yes (H) in *p14_1_2* (mainOliveOil); Less >=4spoon (-HM) in *p14_2_2* (oliveOil); **More** >=2day (H) in *p14_3_2* (vegetables); Less >=3day (-HW) in *p14_4_2* (fruit); **Less** >=1day (-H) in *p14_5_2* (redMeat); **Less** >=1day (-H) in *p14_6_2* (butter); Less >=1day (-H) in *p14_7_2* (gasDrinks); Less >=7glass/week (-H) in *p14_8_2* (wine); **Less** >=3week (-H) in *p14_9_2* (legume); **More** >=2week (H) in *p14_11_2* (commercialBakery); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce); *ge6_12s_2* (intenseWeek) is higher(H); *gehos_2* (homeWorkWeek) is lower(); *getots_2* (totalWeek) is higher(HM); *ge4_5_5a_2* (moderateYear) is higher(); *gehoa_2* (homeWorkYear) is lower();

W-ProtoCaloric : **More** Woman (H) in *sexo* (gender); *EDAD* (Age) is lower(-HW); *altura2* (height) is lower(-HW); *peso2* (weight) is lower(-H); *imc2* (BMI) is lower(-HM); *cintura2* (waist:2) is lower(-H); *pas_esds_fin* (Systolic pressure) is lower(-H); *fc_fin* (Heart Rate) is lower(-H); Less TRUE (-H) in *menopausia_bin* (menopause); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); Less TRUE (-H) in *mdolor2* (NSAID); More TRUE (H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); **Less** TRUE (-H) in *motros2* (other); *gluc3* (glucose) is lower(-H); *cholest1* (cholesterol) is lower(-S); *LDL_ 3* (LDL) is lower(-S); *tryg3* (triglycerides) is lower(-HM); *oxldl3* (Oxidized LDL) is lower(); *isoprostanos_ 3* (F2 α Isoprostanes) is lower(-H); *ifn_ g_ 3* (Interferon-γ) is higher(); *s_ p_ selectin_ 3* (sP-selectin) is higher(S); *s_ cd40l_ 3* (sCD40 Ligand) is higher(S); *oxo_ gg_ 3* (8-Oxoguanine) is lower(-H); *t_ pa_ 3* (Tissue Plasminogen Activator) is lower(-H); More yes (H) in *p14_ 1_ 2* (mainOliveOil); **Less** >=2day (-H) in *p14_ 3_ 2* (vegetables); **More** >=3day (H) in *p14_ 4_ 2* (fruit); More >=1day (HW) in *p14_ 5_ 2* (redMeat); Less >=1day (-H) in *p14_ 6_ 2* (butter); Less >=7glass/week (-H) in *p14_ 8_ 2* (wine); More >=3week (HW) in *p14_ 9_ 2* (legume); **More** >=2week (H) in *p14_ 11_ 2* (commercialBakery); **More** >=3week (H) in *p14_ 12_ 2* (nuts); More yes (H) in *p14_ 13_ 2* (whiteMeat); More >=2week (H) in *p14_ 14_ 2* (sauce); *gehos_ 2* (homeWorkWeek) is higher(SM); *ge6_ 12a_ 2* (intenseYear) is higher(SM); *gehoa_ 2* (homeWorkYear) is higher(SM);

WM-HOO : **More** Woman (H) in *sexo* (gender); *EDAD* (Age) is higher(H); *altura2* (height) is lower(-H); *pas_ esds_ fin* (Systolic pressure) is higher(H); *pad_ esds_ fin* (Diastolic pressure) is higher(H); More ex>5 (S) and less Yes (-H) in *tabaco2* (tobacco); More Divorciado/a (H) and less soltero (-H) in *est_ civ2* (civilState); **More** TRUE (H) in *menopausia_ bin* (menopause); Less TRUE (-H) in *depre* (depression); More TRUE (H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); Less TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); More TRUE (H) in *mhormo2* (hormon); *gluc3* (glucose) is higher(SW); *hdl3* (HDL) is lower(-S); *isoprostanos_ 3* (F2 α Isoprostanes) is (LI); *s_ p_ selectin_ 3* (sP-selectin) is lower(-H); *s_ cd40l_ 3* (sCD40 Ligand) is lower(-H); *pcr3* (C-Reactive Protein) is higher(H); *oxo_ gg_ 3* (8-Oxoguanine) is lower(-H); *t_ pa_ 3* (Tissue Plasminogen Activator) is higher(HM); *il_ 6_ 3* (Interleukin 6) is lower(-SM); *il_ 10_ 3* (Interleukin 10) is lower(); *tyru3* (Tyrosol) is lower(-H); More yes (H) in *p14_ 1_ 2* (mainOliveOil); **More** >=4spoon (H) in *p14_ 2_ 2* (oliveOil); Less >=3day (-S) in *p14_ 4_ 2* (fruit); **Less** >=1day (-H) in *p14_ 5_ 2* (redMeat); More >=1day (HW) in *p14_ 6_ 2* (butter); More >=7glass/week (H) in *p14_ 8_ 2* (wine); **Less** >=3week (-H) in *p14_ 9_ 2* (legume); **More** >=3week (HM) in *p14_ 10_ 2* (fish); Less >=2week (-H) in *p14_ 11_ 2* (commercialBakery); Less >=3week (-H) in *p14_ 12_ 2* (nuts); More yes (H) in *p14_ 13_ 2* (whiteMeat); More >=2week (H) in *p14_ 14_ 2* (sauce); *ge6_ 12s_ 2* (intenseWeek) is lower(-S); *getots_ 2* (totalWeek) is lower(-S); *ge6_ 12a_ 2* (intenseYear) is lower(-S); *getota_ 2* (totalYear) is lower(-S);

WM-HB : **More** Woman (H) in *sexo* (gender); *EDAD* (Age) is higher(H); *imc2* (BMI) is higher(SW); *cintura2* (waist:2) is higher(S); *pas_ esds_ fin* (Systolic pressure) is higher(H); *pad_ esds_ fin* (Diastolic pressure) is higher(H); *fc_ fin* (Heart Rate) is higher(); More ex1-5 (HW) and less Never (-SM) and less Yes (-H) in *tabaco2* (tobacco); More casado/a (SW) and less soltero (-H) in *est_ civ2* (civilState); **More** TRUE (H) in *menopausia_ bin* (menopause); *pertensa* (stressful) is higher(S); Less TRUE (-H) in *depre* (depression); More TRUE (H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); **Less** TRUE (-S) in *mdolor2* (NSAID); Less TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); More TRUE (H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); **Less** TRUE (-H) in *motros2* (other); *gluc3* (glucose) is higher(SW); *cholest1* (cholesterol) is higher(H); *LDL_ 3* (LDL) is higher(H); *hdl3* (HDL) is higher(S); *tryg3* (triglycerides) is higher(HM); *oxldl3* (Oxidized LDL) is higher(H); *isoprostanos_ 3* (F2 α Isoprostanes) is lower(-HM); *mcp_ 1_ 3* (Monocyte Chemotactic Protein-1) is lower(-H); *s_ p_ selectin_ 3* (sP-selectin) is lower(-H); *s_ cd40l_ 3* (sCD40 Ligand) is lower(-H); *pcr3* (C-Reactive Protein) is higher(H); *oxo_ gg_ 3* (8-Oxoguanine) is lower(-H); *t_ pa_ 3* (Tissue Plasminogen Activator) is higher(H); *il_ 6_ 3* (Interleukin 6) is lower(-SM); *il_ 10_ 3* (Interleukin 10) is lower(-S); *tyru3* (Tyrosol) is lower(-H); More yes (H) in *p14_ 1_ 2* (mainOliveOil); **More** >=2day (H) in *p14_ 3_ 2* (vegetables); **Less** >=3day (-HM) in *p14_ 4_ 2* (fruit); **Less** >=1day (-H) in *p14_ 5_ 2* (redMeat); **Less** >=1day (-H) in *p14_ 6_ 2* (butter); **Less** >=1day (-H) in *p14_ 7_ 2* (gas-

Drinks); **Less** >=7glass/week (-H) in *p14_8_2* (wine); **Less** >=3week (-H) in *p14_9_2* (legume); **Less** >=3week (-H) in *p14_10_2* (fish); **Less** >=2week (-H) in *p14_11_2* (commercialBakery); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce); *ge2_3_5s_2* (lightWeek) is higher(HL); *ge4_5_5s_2* (moderateWeek) is (-WI); *ge6_12s_2* (intenseWeek) is lower(-S); *getots_2* (totalWeek) is lower(-S); *ge6_12a_2* (intenseYear) is lower(-S); *getota_2* (totalYear) is lower(-SM);

WM-BasicBak : **More** Woman (H) in *sexo* (gender); *EDAD* (Age) is higher(H); *altura2* (height) is lower(-H); *peso2* (weight) is lower(-HW); *pas_esds_fin* (Systolic pressure) is higher(HM); *fc_fin* (Heart Rate) is lower(-S); More Never (S) and less Yes (-H) in *tabaco2* (tobacco); More casado/a (SW) and less soltero (-H) in *est_civ2* (civilState); **More** TRUE (H) in *menopausia_bin* (menopause); Less TRUE (-H) in *depre* (depression); More TRUE (H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); More TRUE (H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); More TRUE (S) in *mdolor2* (NSAID); Less TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); **Less** TRUE (-H) in *motros2* (other); *cholest1* (cholesterol) is higher(H); *LDL_3* (LDL) is higher(H); *hdl3* (HDL) is higher(S); *tryg3* (triglycerides) is higher(H); *oxldl3* (Oxidized LDL) is higher(H); *isoprostanos_3* (F2 α Isoprostanes) is lower(-HM); *s_p_selectin_3* (sP-selectin) is lower(-H); *s_cd40l_3* (sCD40 Ligand) is lower(-H); *pcr3* (C-Reactive Protein) is higher(H); *oxo_gg_3* (8-Oxoguanine) is lower(-H); *t_pa_3* (Tissue Plasminogen Activator) is lower(-HM); *il_6_3* (Interleukin 6) is higher(); *il_10_3* (Interleukin 10) is higher(S); *tyru3* (Tyrosol) is lower(-H); *ohtyru3* (OHTyrosol) is lower(); *mohtyu3* (MOH_Tyrosol) is lower(); More yes (H) in *p14_1_2* (mainOliveOil); **More** >=2day (H) in *p14_3_2* (vegetables); **Less** >=1day (-H) in *p14_5_2* (redMeat); **Less** >=1day (-H) in *p14_6_2* (butter); **Less** >=1day (-H) in *p14_7_2* (gasDrinks); **Less** >=7glass/week (-H) in *p14_8_2* (wine); **Less** >=3week (-H) in *p14_9_2* (legume); More >=3week () in *p14_10_2* (fish); **More** >=2week (H) in *p14_11_2* (commercialBakery); Less >=3week (-S) in *p14_12_2* (nuts); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce); *ge6_12s_2* (intenseWeek) is lower(-HM); *gehos_2* (homeWorkWeek) is higher(); *getots_2* (totalWeek) is (-LI); *ge4_5_5a_2* (moderateYear) is lower(); *ge6_12a_2* (intenseYear) is lower(-S); *getota_2* (totalYear) is lower(-SM);

WM-ProtoCaloric : **More** Woman (H) in *sexo* (gender); *EDAD* (Age) is higher(H); *altura2* (height) is lower(-H); *imc2* (BMI) is higher(SW); *cintura2* (waist:2) is higher(S); *pas_esds_fin* (Systolic pressure) is higher(H); *pad_esds_fin* (Diastolic pressure) is higher(H); More ex0-1 (S) and less Yes (-H) in *tabaco2* (tobacco); More Separado/a (S) and less soltero (-H) and more Viudo/a (HW) in *est_civ2* (civilState); **More** TRUE (H) in *menopausia_bin* (menopause); *pertensa* (stressful) is lower(); Less TRUE (-H) in *depre* (depression); Less TRUE (-H) in *cancer* (cancer); Less TRUE (-H) in *fractura* (bone fracture); Less TRUE (-H) in *disnea2* (dyspnea); Less TRUE (-H) in *maspirina2* (aspirin); Less TRUE (-S) in *mdolor2* (NSAID); Less TRUE (-H) in *mansiedad2* (anxiolytic); Less TRUE (-H) in *mvitMin2* (vitamin or minerals); Less TRUE (-H) in *mtension2* (heart); Less TRUE (-H) in *mhormo2* (hormon); More TRUE (H) in *motros2* (other); *gluc3* (glucose) is lower(); *cholest1* (cholesterol) is higher(H); *LDL_3* (LDL) is higher(H); *tryg3* (triglycerides) is higher(HM); *oxldl3* (Oxidized LDL) is higher(H); *mcp_1_3* (Monocyte Chemotactic Protein-1) is lower(-H); *s_p_selectin_3* (sP-selectin) is lower(-H); *s_cd40l_3* (sCD40 Ligand) is lower(-H); *oxo_gg_3* (8-Oxoguanine) is lower(-H); *t_pa_3* (Tissue Plasminogen Activator) is higher(H); *il_10_3* (Interleukin 10) is lower(-S); *tyru3* (Tyrosol) is lower(-HM); More yes (H) in *p14_1_2* (mainOliveOil); **Less** >=2day (-H) in *p14_3_2* (vegetables); **More** >=3day (H) in *p14_4_2* (fruit); **Less** >=1day (-H) in *p14_5_2* (redMeat); **Less** >=1day (-H) in *p14_6_2* (butter); **Less** >=1day (-S) in *p14_7_2* (gasDrinks); **Less** >=7glass/week (-H) in *p14_8_2* (wine); More >=3week (H) in *p14_9_2* (legume); **More** >=3week (S) in *p14_10_2* (fish); **More** >=2week (H) in *p14_11_2* (commercialBakery); **More** >=3week (H) in *p14_12_2* (nuts); More yes (H) in *p14_13_2* (whiteMeat); More >=2week (H) in *p14_14_2* (sauce); *ge2_3_5s_2* (lightWeek) is higher(); *ge6_12s_2* (intenseWeek) is lower(-SM); *ge2_3_5a_2* (lightYear) is higher(S); *ge4_5_5a_2* (moderateYear) is higher(SM); *getota_2* (totalYear) is higher();