ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis: Data Science and Marketing Analytics

Identifying Chilean Dietary Patterns, Key Groups and Sociodemographic Drivers in Aims to Meet The Sustainable Development Goal 12

María Jesús Bascuñán Moraga

July 3rd, 2020

Student ID: 523367

Supervisor: Ilker Birbil

Second Assessor: Stacey Malek

Abstract

Worldwide, food is one the categories of products with the most important environmental footprints due to an excessive consumption that is driving an indiscriminate over-exploitation of natural resources. In the quest of complying with the Sustainable Development Goals, countries are developing strategies to raise awareness about responsible consumption, among them Chile, the focus of this thesis. Thus, it has become relevant to identify key groups with certain dietary patterns in order to develop public policies to target this segments. Data comprising food consumption, regional characteristics and food's environmental footprints was gathered and by means of Principal Components Analysis applied over foods frequency intake, five dietary patterns were identified. Subsequently, clustering method k-Prototypes is applied over the dietary patterns in conjunction with footprint and healthy eating indices, discovering six clusters of people. Finally, supervised methods are performed in order to understand the importance and associations between sociodemographic variables and the clusters. It is determined that age, gender and socioeconomic level, with different intensities across clusters, are the main features that help classifying an individual. Furthermore, geographic features, such as the Region or rurality of the area, become relevant when identifying certain clusters.

Acknowledgements

First, I wish to thank my assessor, Prof.dr. Ilker Birbil, for his guidance and advice during this project. I would also like to thank my family and friends for their constant support during my experience and study in Erasmus University. Finally, I feel very grateful for the help provided by my classmates from University of Chile in the process of identifying challenges for Chile and retrieving data related to the matter, everything in aims to build something useful for the community and government.

Key Words

Dietary Patterns, Sustainability, Environmental Footprint, Principal Components Analysis, k-Prototypes, Classification, Multinomial Logistic Regression, Random Forest.

Content

List of Tables

List of Figures

1. Introduction

In 2015, the United Nations (UN) Member States committed to 17 Sustainable Development Goals (SDGs) to achieve a sustainable future for everyone. The SDGs take part of the 2030 Agenda for Sustainable Development. A 15-year plan was set out to achieve the goals, which are interrelated and comprise global challenges faced every day, such as poverty, health, inequality, climate change, environmental degradation, prosperity, peace, and justice. New public policies about diverse topics have been already implemented, others are being developed. New studies, articles, and models are published every day, all with the common objective of working towards a better future, trying to understand the current situation, identify patterns and predict possible outcomes given a plethora of factors. And new technologies are arising to keep track of the environmental situation. In simple words, progress is being made, however, not at the required speed or scale (United Nations, 2020. Sustainable Development Agenda).

Even though most of the success of meeting the SDGs relies upon a handful of experts and policy-makers in each country and region, people still have a role to play when it comes to achieving some of the SDGs. This thesis aims to put its focus on the SDG 12: "Ensure sustainable consumption and production patterns", and more specifically on the goal 12.8: "By 2030, ensure that people around the world have relevant information and knowledge for sustainable development and lifestyles in harmony with nature." Through awareness-raising on sustainable consumption and lifestyles, this goal intends to engage people in this quest for a better life for them and the planet. The scope of this study will be specifically sustainable food consumption.

Food is one of the key elements that need to be assessed from a variety of perspectives, for the reason that it has an impact on several aspects related to sustainability, therefore being relevant for different SDGs. Why is food important? Firstly, food is directly related to nutrition, thus also related to the health of an individual. Secondly, what people demand eating *ipso facto* affects the food industry economics and production for the next

seasons. Natural resources are being excessively exploited due to the irrational use of food, which also leads to the unmet nutritional needs of peoples. Land and marine environment degradation, overfishing, lower soil fertility, and unsustainable use of water are the results of this over-exploitation in means to supply food which in return is also lessening the ability of the natural resource base to keep providing aliments. Thirdly, food consumption goes hand in hand with food waste, where the losses are borne along the entire food chain, which has an impact on economic, social, and environmental dimensions. Altogether, the process of food production implies leaving a carbon footprint in different stages, namely life-cycle, from sowing the seeds and feeding the cattle to transporting the products and having them on the table, accounting for 22% of total greenhouse gas emissions and 30% of the world's total energy consumption (United Nations, 2020. Goal 12). These last two points are the ones most strongly related to SDG 12 because households influence the impacts of food production life-cycle through their dietary choices and habits. Consequently, and as mentioned previously, the environment suffers from food-related energy consumption and waste generation.

Several kinds of research have been performed aiming to identify diets that result in being more sustainable than the ones with significant consumption of meats and dairy as daily components of the meals. Diets low in the later mentioned products and high in nuts, seeds, and legumes as a source of protein are more environmentally friendly, and recent studies prove that plant-based diets have further health benefits than omnivorous diets.

There is a myriad of factors that strongly influence dietary choices and patterns, beyond personal decisions (BJM, 2018). Personal preferences, gender, health condition, education, culture, and even cooking skills can determine dietary habits. Psychological elements such as attitudes, incentives, and motivation also influence food choices. Even a mother's diet during pregnancy and feeding during upbringing have an effect. Household's mores, sleeping habits, social class, norms, and race are also relevant elements. Furthermore, commercial pressures such as marketing and new trends on social media also influence food choices. However, well-studied, thoughtful, and data-driven public policies can help

improve households' diets, health, and overall wellbeing. The medical journal BJM in 2018 published a study on government's policy interventions, voluntary and mandatory for the public, that better nutrition and prove to be effective. In this case, people's awareness about sustainable consumption, which is a major factor in the quest to meet the SDGs, should be boosted by public policies developed by governments.

Chile is one of the UN State Members that adopted the SDG and thus committed to the 2030 Agenda. Chile has a population of over 19 million inhabitants, a GDP of 227,1 thousand of millions of dollars, ranking 41st worldwide, and 15,3 thousand dollars per capita, ranking 53rd worldwide. It was the first Latin American country to be part of the Organization for Economic Cooperation and Development (OECD) and it is expected for it to be a developed country by 2025.

Like all the other UN State Members, Chile has been working in aims to meet the goals settled, having as the main focus to improve the quality of life of people (reducing poverty, bettering the health system, reducing inequality, better education) and the use of clean energies. However, the fast-growing economy deflects the efforts to protect the environment on a large scale, despite recent public policies and initiatives that have arisen from local communities or groups of activists. Introduction of regulations that help to have broader control over the use of natural resources and the production system footprints are necessary, however, such changes are slow processes and the effectiveness of its impacts would be seen in the long term, which is something the planet cannot afford. So for now, other paths or a faster speed when implementing public policies will have to serve as a solution to meet the SDG.

Food consumption in Chile represents 18,7% of the total spending of a household on average (Marchetti 2018, Emol), thus it is an important source of carbon emissions and other environmental footprints. In this way, is it possible to shift food consumption behaviours in such a way that they become more sustainable? Which groups of the population should the government target first? What sociodemographic and regional

factors influence these diets? Such questions have not been yet answered and could be summarized as the main research question for this thesis:

Which are relevant groups and the sociodemographic variables that affect them on which the Chilean government should focus to comply with the Sustainable Development Goal 12?

Coming into effect in June 2016, the public sector of Chile developed the Food Law, pioneering in public policies regarding nutrition. This new law intended to reduce children's obesity by educating about nutrition and promoting physical activity and by labelling products warning "high in"' calories, saturated fats, sugar, and sodium. Advertisement of products with any kind of "high in" label targeting children younger than 14 years old was forbidden and the sale of these products was prohibited in schools. In 2009-2010 Chile had the highest proportion of overweight population among the countries in the region and members of the OECD. This is further enhanced among people with low educational level and low and middle socio-economic level. If changes were not made, the resources allocated on the health system to attend obesity-related health issues would increase from 0.5% of GDP in 2006 to 1.6% in 2030 (Food and Agricultural Organization of the United Nations, 2017).

During the XVIII Latin American Nutrition Congress held in 2018 in Mexico, the Chilean delegation presented the successful results of the Food Law, showing how public policies can have a significant impact not only over people's conduct but also over that of producers. For instance, the consumption of carbonated drinks dropped by 25%, breakfast cereals by 36%, and package desserts by 17%, whereas the intake in highly nutritive food presented an upsurge of 9%. The new law also resulted in the reformulation of many food items from different categories, lowering the addition of sodium, sugars, and saturated fats (Institute of Nutrition and Food Technology of University of Chile, 2019).

The Government of Chile already launched a national program for the years 2017 - 2022, designed by the interministerial Committee for Sustainable Consumption and Production that aims to serve as a guide in matters of sustainable consumption and production

(Ministry of Environment of Chile, 2017). One of the axes of this program corresponds to actions focused on providing more information for the consumer, axis projected to be fully developed by 2020, which includes the certification and labeling of products, the development of indicators for products and services, communication of information and diffusion of sustainable trends, and, lastly, education. Another axis is concentrated in sustainable lifestyles, which includes the diagnosis of the current lifestyles in Chile and the identification of target groups, designing and implementing a communication strategy aiming to change habits and adopt sustainable lifestyles, and the definition of key indicators. It is important to mention that this program of plans of action is merely a framework, and the definitive implementation depends on each work team assigned per axis; furthermore, the compilation of initiatives was not exhaustive nor been audited.

Based on the above, and with the understanding that changes in diets imply changes of behaviours on the part of the consumers, one of the outputs of this study are some recommendations on where the government of Chile should focus its efforts on generating awareness regarding sustainable diets, in order to comply with SDG 12. Discovering food intake patterns enables the identification of key groups that are the furthest from presenting sustainable diets, determining which foods should be increased or diminished in consumption and what groups to target. While finding food intake patterns, a better understanding of how household characteristics and Region climatic and socioeconomic features influence these patterns can also be derived. Even though the model is focused on Chile, it can be generalized and scaled to become useful for other countries that want to look for more edges to tackle the SDG 12.

This thesis comprises 6 sections. In Section 2, a literature review is presented, mostly focused on methods to find dietary patterns and how to measure and assess the sustainability and healthiness of a diet according to different approaches. In Section 3, the data on which the study is based is described. In Section 4, the methodology is explained. Section 5 presents the results and Section 6 exposes the main conclusions and recommendations for stakeholders and future investigations.

2. Literature Review

Due to the nature of this research that crosses different fields of study, the literature review has three main axes. First, in Section 2.1 methods to identify dietary patterns retrieved from several papers are presented. Second, a popular procedure on how to assess the environmental impact of foods can be found in Section 2.2. Lastly, the relevant indicators of healthy eating available nowadays take part in Section 2.3. The techniques and indicators exposed throughout this review served as guidelines for the choices and methodologies applied in this study.

2.1 Identifying Dietary Patterns

Dietary patterns are understood as measures of the usual consumption of food combinations in individuals and groups (Tucker, 2010). Most of the studies that base its findings on dietary patterns analyses are motivated by the goal of unraveling and comprehending the relationships between food intake and health outcomes, such as nutrient deficiency diseases or chronic conditions, to provide insightful recommendations for health institutions and governments. Some studies are focused on large groups of populations with diverse characteristics and others focus on specific cohorts, like pregnant women from a certain country or indigenous ethnicities from a specific region. In most of the cases, cultural and socio-demographic aspects are also considered when trying to provide a further understanding of food consumption behaviours and its diverse consequences for the human body. Nevertheless, studies usually focus on the characteristics of the subject of study and not in the characteristics of the area to which they belong to (such as average temperature, land use, rurality, and so on).

Most studies are based on individual-level surveys, indexes, or databases in general. These data sets comprise the food intake of individuals or households in different formats. For instance, food intake in some cases is captured in terms of frequency, in terms of the amount consumed (grams, for example), in terms of intake of nutrients or percentage of energy intake. In several instances, food items are grouped beforehand due to the vast

amount of variety (for example, spaghetti and gnocchi may fall under the same category, namely pasta, or lentils and beans may be grouped as legumes). Regarding the time frame, food intake can be measured daily or monthly, and some studies include both. However, it is worth mentioning that because diets vary from day to day, the analyses usually consider broader periods, such as monthly average food intake, thus deeming databases with daily food consumption usually leads to poor identification of truthful patterns and misclassification issues.

There are two main approaches on how to define food groups or dietary patterns. The first one is based on the recommendations of health public institutions, health and nutrition journals, and criteria of experts, among others, resulting in "pre-defined" patterns that are afterward used as input in aims to identify relationships with diseases or socio-demographic variables. The second approach is algorithm-based or multivariate statistical methods, applied over the food intake from collected dietary information. More specifically, the dietary patterns are identified by means of clustering techniques and by the reduction of dimensions, where the original dimensions correspond to each food item, food group, or nutrient. Furthermore, dimension reduction usually precedes the application of clustering techniques. Thus, two algorithms are mainly used, Cluster Analysis and Factor Analysis, such as Principal Component Analysis (PCA). Some authors further elaborated on this last topic, but also Mixture Models (MMs) have been used in explorative nutritional epidemiology. Moreover, Latent Class Analysis (LCA) can identify subgroups within a population, and then the resulting groups could be used for regression analysis, resulting in a MM. Dietary prototypes have also been discovered by means of LCA over medium and large size samples (Hammerling et al., 2014).

Nishi et al. (2017), by using pre-defined patterns in a multilevel regression modelling, discovered that individuals from households with low or middle income consume more staple food items, like cereals, in comparison to high-income households, which consume more vegetables, fruit, and fish. Furthermore, individuals from middle-income households present a higher energy intake when contrasted with individuals from low-income

households. Another study performed in the US, also utilizing pre-defined dietary patterns, revealed a positive association between having frequent family meals and adolescents with high socioeconomic status. Family meals, in turn, are related to high energy intake and consumption of fruits, vegetables, grains, and calcium-rich foods (Neumark-Sztainer et al., 2003).

Padmadas, Dias, and Willekens (2007) used LCA over a data set comprising intake frequencies of selected food items from 90,185 women in India, intending to examine different dietary intake patterns. Through this approach, clusters are formed, and then the cluster memberships are treated as the dependent variable in a multinomial logistic regression (MLR) where the predictors are demographic, spatial, socio-economic, and cultural characteristics of the individual. This aimed to explain associations among the observed variables of the women. The cluster labels were set according to how mixed was the diet between vegetarian and non-vegetarian dietary intakes, namely very high mixed-diet, high mixed-diet, moderate mixed-diet, low mixed-diet, and very low mixed-diet. One of the findings, for instance, was that very high or high mixed-diet patterns are mostly found in southern and a few north-eastern states in India.

Ambrosini et al. (2008) researched dietary patterns and prostate cancer risk among men in Western Australia using a population-based case-control study comprising 685 individuals, and similar to the study presented in the previous paragraph, they applied dimension reduction, but in this case by means of PCA, discovering three dietary patterns, namely "vegetable" (intensive in nearly all vegetables, plus honey, apples, and jam), "Western" (intensive in carbs, such as white bread, cakes, and potato crisps, eggs, red meat, processed meat, and beer) and "health-conscious" (intensive in white meat, rice, legumes, tofu, beans, nuts, yogurt, and wine). Later, logistic regression models were used to estimate prostate cancer risk, where the "Western" pattern conferred a higher risk of prostate cancer. Additionally, associations between characteristics of the subjects and the diets were found, for instance, a high score in the "Western" pattern is related with a

larger portion of smokers and smaller portion of men that exercise regularly, opposed to the observed for the "vegetable" and "health-conscious" patterns.

Lazaroua et al. (2011), based on a subset of 634 observations extracted from a nation-wide, cross-sectional survey of children from Cyprus comprising consumption frequency (measured in terms of a four point-Likert like scale) of a variety of food items and socio-demographic characteristics, first performed classification by decision tree (DT) induction, per gender, thus different thresholds for intake frequency of selected food groups came up as a result of each splitting node in order to classify the individual in different weight status categories. Several rules were derived from the classification trees and were ordered and summarized. Then they also performed PCA, coming up with four factors which afterward were used as independent variables in a MLR to understand the relationship between these variables and status of normal weight or overweight/obese. The main discovery was that boys appear to present a higher intake of the food items that comprise factors described as intensive in fried food, junk food, and sweets, and intensive in dairy products and protein food. However, no significant relationship was found regarding weight or other variables such as place of living (rural/urban), ethnicity, and socio-economic level.

Kastorini et al. (2013), with the aims of comparing the accuracy of *a priori* (from MedDietScore) and *a posteriori* (employing PCA) dietary patterns in the prediction of acute coronary syndrome and ischemic stroke, they enrolled 1,000 participants. Using six classification algorithms, they managed to model the relationships between the conditions mentioned and the dietary patterns identified (five principal components or patterns in total). The classification algorithms corresponded to MLR, Naïve Bayes, DT, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), Artificial Neural Networks (ANN), and Support Vector Machines. In this study, the best ones were RIPPER and MLR (based on the C-statistic), however, both presented similar classification accuracies, thus the choice depends on the application.

Hammerling et al. (2014), based on two food surveys from Sweden and Denmark, used a top-down multi-branching hierarchical clustering (HC) algorithm to identify eight dietary prototypes. Subsequently, the Danish and Swedish databases were subjected to variable selection and training of the Nearest Shrunken Centroid (NSC) classifier, thereby enabling multivariate pattern recognition. The classifiers were built to discriminate between low and high consumers of each food category studied, based on their dietary patterns (excluding the food category considered). In simpler words, for example, a Candy classification model was built to classify High/Low consumers of Candy based on the consumption pattern across the remaining foods. This procedure was conducted to unveil associations with the respective foods, not readily disclosed by any exploratory methodology. For instance, consumption patterns for Danish and Swedish children are quite distinct, revealing how cultural differences may affect diets, however, some food items, such as meat, potatoes, bread, milk and poultry are linked to particularly different food patterns.

In 2019, Huseinovic et al. published an article in the Nutrition Journal where they presented the identification and analysis of changes in food intake patterns in Sweden based on a Diet Database with almost 50,000 observations, linking sociodemographic variables and personal characteristics that affect these patterns by using LCA. The data was split into men and women and between two periods of time, 2000-2007 and 2008-2016, thus resulting in four sections. The patterns were constructed using frequency (adjusted by 1,000 calorie energy intake) rather than the amount of intake, resulting in four clusters for each of the four sections. The clusters were named after the food groups with the highest intake frequency. Subsequently, conditional means were observed but no statistical test that reveals significance was applied. About the findings, focusing in the case of women in the time frame of 2000-2007, by way of illustration, the most common cluster is "High-fat dairy, white bread, sugar/jam and cookies", followed by "Fruit, high-fiber bread and low-fat milk", "Bacon/sausage and fast food", and "Pulses and tea." More specifically, it is shown that women belonging to the fourth cluster are more physically

active and have a higher educational level. Furthermore, the second and fourth clusters are characterized by consuming much healthier meals when compared to the other clusters among women, having fewer smokers.

Hearty and Gibney (2008), for example, used ANN and DT to predict quintiles of the Healthy Eating Index based combinations based on the North-South Ireland Food Consumption Survey 1997-1999, which they manage to do with high accuracy.

Most of the studies mentioned in this section had a health-related objective, however, some studies place its focus on environment-related goals. For instance, in 2018, in the Climatic Change journal was published a study by Biesbroek et al. that aimed to identify data-driven healthy dietary patterns that benefit the environment; benefit for the environment in the sense that produce less greenhouse gas emissions, which were calculated using life cycle analysis. Afterward, the patterns were identified with Reduced Rank Regression.

In spite of the fact that the presented studies did provide interesting outputs, some limitations need to be taken into consideration when applying those methods. For instance, when it comes to supervised methods, DT can easily overfit the sample if not pruned correctly; nowadays Random Forest and Boosting could provide more accuracy and less overfitting. On the other hand, the identified latent classes or components derived from PCA or LCA, adopt a rather abstract nature in the sense that it is not evident what combinations of foods comprise each new variable. It has been shown that applying PCA particularly over food items to obtain dietary patterns often results in principal components that together do not explain more than 10% to 40% of the variability of the original data, when it is expected to be over 80% in general applications. For example, in the study by Thorpe et al. (2016), where they performed PCA over 52 food items and then k-Means to create clusters, the four principal components identified for men presented a total variance explained of 22.6% and the two components for women a total of 14.3%. Furthermore, Tseng et al. (2004) found three components with a total variance

explained of around 11% and McCann et al. (2001) showed that variance explained can upsurge if food items are grouped, increasing from 8% with 168 items to 17% with 36 items, however, this did not improve the prediction of the disease under study. This low variance explained is due to the high dimensionality of the data. In general, the choice of components is driven by total variance explained, by examining the resulting scree plot, by selecting components with eigenvalues strictly larger than one, and by the expert criteria of the authors, that usually look for reasonable and interpretable patterns.

When it comes to Cluster Analysis, even though cross-validation can be used to examine the resulting clusters, by looking at within-cluster variance or by scree plots, there is not an ultimate technique to define the appropriate number of clusters. This means that an important part of the process of determining the number of groups relies on the author's criteria (Devlin et al., 2012). Also, due to the fact that it would be an enormous amount of work to gather information about every exact food item consumed, in most studies the food items are grouped or coded beforehand, subsequently influencing the methods applied. Interesting patterns associated with cultural aspects of different subpopulations could be overlooked. Thus, data pre-processing plays an important role when trying to unveil insights of the population's diet.

2.2 Measuring the Sustainability of a Diet

As introduced in the first section, the aim of this study is to uncover dietary patterns in order to understand how far these are from a sustainable diet. Thus, a definition and measurement of what a sustainable diet is, is necessary.

To understand the impact of different food items, and therefore of different diets, a rather popular approach can be performed, named product Life-Cycle Assessment (PLCA), as done by Biesbroek et al (2018). PLCA is a holistic view of environmental interactions occurring during production, transportation, retail, and consumption of a product. It considers the resources used and environmental releases, starting with the extraction of raw materials and the disposal of the final products. In the case of food, it also comprises

preparation and losses. PLCA allows us to compare food items in terms of eco-friendliness, for example. This technique has several steps, such as: choosing a functional unit (CO_2 emissions, e.g.); delineate the goal and scope of the study; compiling a catalog of pertinent energy and material inputs and environmental releases; evaluating the potential environmental impacts; and, lastly, interpreting the results to help interested individuals and organizations to make more insightful decisions (Curran, 2008).

An application of this method can be found in the article produced by Salas and Castellani (2019), where it was discovered that the food products that have the most impactful footprint are animal-based (meat and dairy) and beverages, and from these, the main processes that contribute are the ones associated with animal feeding.

Several publications and studies address the enormous footprint associated with animal-based products, being accountable for about 18% of greenhouse emissions, thus a reduction of its consumption or even completely moving towards plant-based protein foods would have an important impact over land use. Furthermore, more recent studies unveil some negative repercussions that animal-based products have over people's health, which in turn has given on to revising dietary recommendations on part of health institutions (Stehfest et al. 2009).

Kramer et al. (2017), based on the Dutch National Food Consumption Survey 2007-2010, tried to optimize the Dutch diet so it would meet the recommended nutrients' requirements and goals over environmental indicators. This was performed by means of linear programming using Optimeal (Blonk Consultants, Gouda, the Netherlands). Optimeal is a tool that by means of optimization intends to find diets both healthy and sustainable. One of their main conclusions is that a reduction in meat consumption is the most fruitful way of lowering the environmental impact of a diet, although some nutrients may reach critical levels when taking this measure, such as vitamin B12, Calcium, and Iron. It was also found that sociodemographic variables, such as age and gender, also

affect the environmental impact of the diet due to the distinctive nutritional requirements. This sheds light on the need to tailor optimal diets for each target group.

Another study performed by González, Frostell, and Carlsson-Kanyama (2011) shows that plant-based proteins are more efficient in terms of emission of greenhouse gases, as opposed to animal-based proteins. The aim of this study was to evaluate how food items consumed by households in Sweden impact climate change, measured in terms of protein efficiency, i.e., "the protein obtained per kg of GHG emitted in the production and transport of the food up to the wholesale point."

To assist companies and institutions in measuring the environmental impact of the products they produce, in Chile in 2015 was launched the Food Ecobase by Fundación Chile, a publicly available calculator in Excel format that delivers the environmental footprint when it comes to carbon, water, and energy, of 16 products from the agrifood and export wines sectors, based on PLCA, such as apples, table grape, salmon, avocado, milk, chicken, among others. This tool was built based on 150 sources and it also includes a national benchmark (UN Global Compact, 2015).

2.3 Measuring a Healthy Diet

Technically, a generalized low consumption of food seems to be the most sustainable diet, however, it is intended to find a middle ground between sustainability and healthiness. For a diet to be nutritionally balanced it must provide the energy and nutrients in adequate quantities and sufficiency to cover the nutritional needs of individuals. Diets are key when it comes to preventing deficiencies and potential diseases. Moderate consumption and variety in the ingested food items are also core recommendations (Ortega & Requejo, 2000).

Several manners of measuring healthy diets have been developed over the years. Usually, each countries' Health Ministry or National Health Service (or analogous institutions) provides dietary recommendations over which health indices are developed. For instance, in Chile, the recommendations by the Ministry of Health are simple, intending to be easily

understandable by the people. The recommendations of food intake are the following: 1- Three portions of dairy per day, equivalent to 600 ml, 2- Five portions of fruits or greens and vegetables (raw or cooked) per day, equivalent to 400 gr, 3- Leguminous (beans, chickpeas, lentils, and peas) two times per week, 4- Fish two times per week and 5- 1,5 lt of water (excluding soft drinks, soda, and soup) (Ministry of Health of Chile, 2013).

From these recommendations and for the National Food Consumption Survey 2010 in Chile, an index named Global Index (GI) was derived, which has three possible outcomes according to the compliance of each recommendation. "Satisfactory compliance" corresponds to complying with at least three out of the five recommendations established by the MINSAL, "partial compliance" corresponds to the compliance of one or two out the five guidelines and "no compliance" when none of the guidelines is followed (University of Chile, 2014).

One of the most popular measurements in the western world is the Healthy Eating Index (HEI), originally developed in the USA in 1995, and now with an updated version from 2015, which is an indicator that attempts to assess how well the food intake aligns with key recommendations of the Dietary Guidelines for Americans (DGA), provided by the U.S. Department of Health and Human Services and U.S. Department of Agriculture (2015). These recommendations aim to help to prevent or reduce the risk of developing diet-related chronic diseases (Krebs-Smith et al., 2018). The HEI, which uses a scoring system that ranged between 0 and 100 points, has been used to describe the US diet quality of its entire population but also of sub-populations. It has also been used as the foundation to find associations between diets and health outcomes in the US. Furthermore, some countries have based their own healthy eating index on the HEI.

Spain is one of the countries in Europe regarded as having what is known as a Mediterranean diet, characterized by a high intake of olive oil, nuts, cheese, fruits, vegetables, pasta, rice, fish and white meats. Spain has the Healthy Eating Index for the Spanish population (HEIS), which introduces some modifications over the HEI created by

the USA to make it applicable to the situation in the country. The HEIS also has a scoring system from 0 to 100 points, but is frequency-based, in the sense that the scores depend on the number of times a certain food item is eaten on a weekly basis, rather than density-based as the HEI from the US. Another difference is that the components, in this case, are actually mostly food groups and then the diversity of the diet, more specifically, the variables are: 1- Cereals and derivatives (like pasta, bread, potatoes, and rice), 2- Greens and Vegetables, 3- Fruits, 4- Dairy and derivatives (like cheese and yogurt), 5- Meat (including eggs), 6- Legumes, 7- Sausages and Cold Cuts, 8- Sweets, 9- Soft drinks with sugar and 10- Diet variety. These variables are categorized according to the recommended intake in a certain period, and it is as follows: the 1st, 2nd, 3rd, and 4th variables represent the food groups to be consumed daily, the 5th and 6th variables correspond to food groups of weekly intake, 7th, 8th, and 9th should be consumed occasionally, and the 10th variable, diet variety, represents the main objective of a healthy diet (see Appendix A). The final classification of the diet of individuals or households according to the scores is "healthy" if it scores over 80 points, between 50 and 79 points "requires changes" and below 50 is "unhealthy" (Norte Navarro and Ortiz Moncada, 2011).

Both indices, the HEI and the HEIS, have in common that they suggest that high intakes of cereals, fruits, and greens are beneficial for health, as well as dairy, especially cheese and yogurt. Nevertheless, some differences appear, for instance, the Spanish version of the index is more explicit when it comes to setting a strict limit score for meat intake, and even more strict with cold cuts, whereas the American version just refers to a suggested amount of proteins and then limits on sodium or fatty acids, components that can be found in meat and cold cuts. When these indices are contrasted with the GI of Chile, it can be noticed that the latter does not give any specific guidelines when it comes to meat consumption, potentially leading to differences when assessing the healthiness of the diet of the population.

3. Data Description

In Figure 3.0.1 the information sources and its treatment are depicted in a flowchart to facilitate the understanding. Three types of sources of information are used in this study. The main one is the National Food Consumption Survey (NFCS) 2010 of Chile, from which individuals' food consumption and sociodemographic variables are obtained, described in Section 3.1. Then, Regional features were extracted from different sources in order to complement the main dataset, presented in section 3.2. Additionally, information about the environmental footprint of food groups and food items was gathered to construct footprint indices, further explained in Section 3.3.

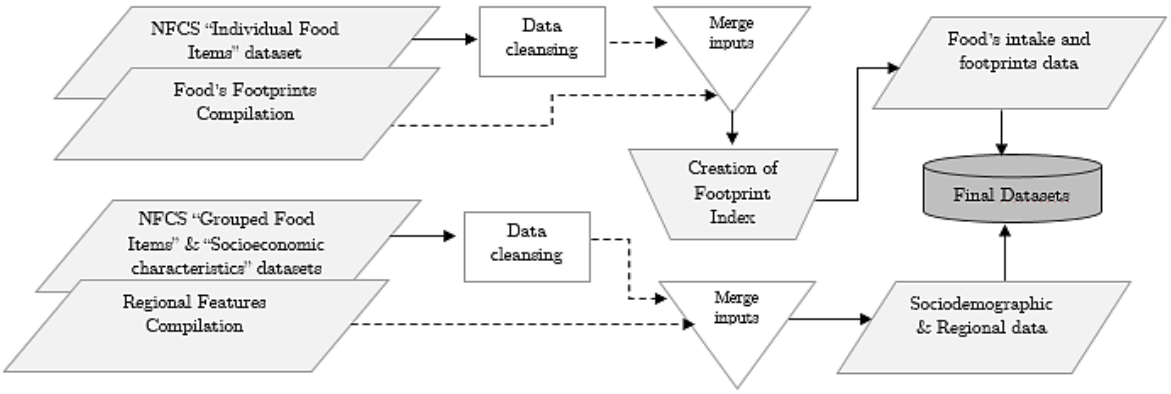

Figure 3.0.1: Flowchart of information data sources and pre-processing

3.1 National Food Consumption Survey

The NFCS 2010 is a socio-economic survey applied over 4,920 individuals of selected house-holds, carried out by the University of Chile between November 2010 and January 2011 in Chile. The results provide information on consumption, eating habits, and nutritional situation with national representativeness (through an Expansion Factor, further explained below). The information reflects the monthly consumption of individuals and also the consumption of the last 24 hours. However, this study is based on the monthly consumption, due to the fact that daily food consumption does not necessarily represent a habit, an issue further addressed in Section 2.1.

There are three databases resulting from this survey that are used in this study: "Individual Food Items", "Grouped Food Items" and "Socioeconomic characteristics." The "Individual Food Items" database comprises the consumption of 469 food items at an individual level, indicating the amount consumed in terms of grams or millilitres and the frequency, both monthly, which is the main source to identify dietary patterns. Selected additional information from the "Grouped Food Items" and "Socioeconomic characteristics" datasets is obtained at an individual-level and a household-level. The variables at an individual-level are Folio (unique identifier), age, gender, Expansion Factor, Body Mass Index (BMS), and two healthy eating indices, namely Global Index of Chile and Healthy Eating Index of Spain. The variables at a household level are socioeconomic level, the number of members of the household, area of residence (urban or rural), city, region, and macro zone, corresponding to the aggregation of regions per zone (Figure 3.1.1). Overall, the final database comprises 19 variables in total (see Appendix B for more details).

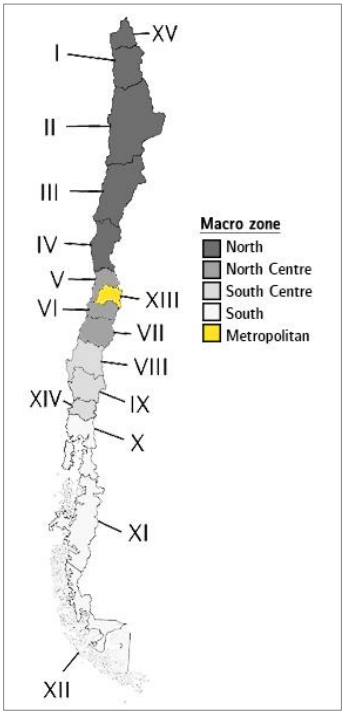

Figure 3.1.1: Regions of Chile and its corresponding macro zones

It is important to mention that the NFCS has a complex sample design, where each observation represents a number of units in the population, captured by the variable called Expansion Factor. Thus, when the expansion factors of each observation are added, it is equivalent to the population, 15,658,607 individuals in this case. The Expansion Factor depends on the sampling design given by the probability of selection of people according to the communes, sector, and members of the household. This variable becomes relevant when optimizing classification models.

Data cleansing was performed in order to work exclusively with observations with data in all the variables. The final database comprises 4,663 individuals from 111 communes and 15 regions of Chile, divided into 2,867 women and 1,796 men. The average age is 41.3 years old, however, there is also a high frequency of people younger than 25 years old (see Appendix C Figure C.1). When checking the distribution of the age groups according to the variable AGE_GROUP, it is worth mentioning that the category "<6" presents zero cases and the categories 5 and 6 are the ones with the highest frequencies. The most frequent socioeconomic levels are medium-low (4) and medium (5), where the first one represents over 35% of the sample (see Appendix C Figure C.2). Around 40% of the individuals are from the Metropolitan macro zone, also observed in the regions' distribution (see Appendix C Figure C.3 and Figure C.4), followed by the macro zone North Centre (20%), South Centre (18%), South (11%) and North (11%).

Additionally, some superficial relationships can be seen when observing the distribution of individuals across combined variables. For instance, the socioeconomic levels vary according to each macro zone and this is even enhanced per region. For instance, it can be said that the South macro zone is the one with the lowest socioeconomic level (see Appendix C Figure C.5).

To facilitate the description of food intake, the food items are grouped in two levels, level 1 is more general, comprising 22 categories, and level 2, with 68 products, which is the

one used to construct the dietary patterns. For instance, in level 1, all types of meat belong to the group meat and all types of fruit belong to the group fruits, whereas in level 2 there is a category for each type of meat, namely beef, lamb, pork, chicken and processed meat, and for each type of fruit, such as citrics, berries, and grapes, other fresh fruits, dried fruits, canned fruits, and so on (see Appendix C Table C.1).

When observing food at an aggregation level 1, the food groups declared to be eaten at least once a month by more than 95% of the individuals are vegetables, oils and butter, meat, bread and fruits, rice, milk and cheese, and sugar (see Appendix C Figure C.6). On the other hand, foods such as cereals, coffee, nuts and seeds, liquors, and soy derivatives are consumed by less than 50% of the people. Regarding the frequency intake, oils and butter, sugar and sweeteners, bread and vegetables are consumed more than 20 times per month on average, thus more than 5 times per week, whereas the least consumed items are fish and seafood, legumes, liquors and nuts, and seeds, with a monthly frequency below 5 (see Figure 3.1.2).

To provide more detail about the specific items most often eaten by Chileans, food at an aggregation level 2 is observed (see Appendix C Figure C.7). White bread, tea, root vegetables, sugar, skim milk, fruit vegetables, leaf vegetables, juices and butter, and margarine are at the top of the list, with an average frequency equal or above 15 times per month. These items are followed by sunflower oil, fresh fruits, other edible oils, potatoes, and soft drinks. With even a lower consumption (around five to eight times per month) poultry, whole milk, pasta, citric fruits, beef, sweet pastries, and berries and grapes can be found.

A closer look is taken to key food groups in a person's diet. Among the meat group, the one with the highest frequency is processed meats, with an average of more than eight times per month, followed by poultry, beef, pork, lamb, and other meats. About the vegetable group, root vegetables are consumed the most often, over five times per week,

followed by fruit vegetables and leaf vegetables consumed around three times per week. Regarding fruits, the category fresh fruits is the one with the highest frequency, which is something expected, because it not only comprises rather common fruits, but also the most affordable fruits, namely apples, bananas, pears, and so on.

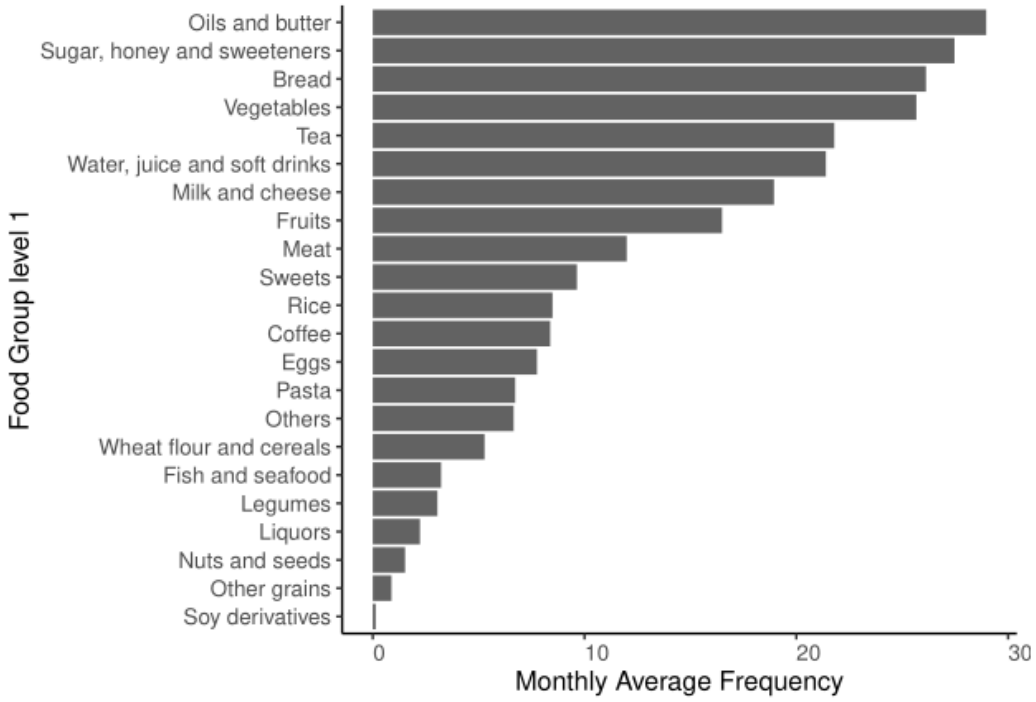

Figure 3.1.2: Intake frequency of grouped food (aggregation level 1)

Concerning health-related variables, the average BMS is 27.0 and the distribution is mostly concentrated around the average. According to the GI of Chile, 12% of the individuals present a "satisfactory compliance", 24% "partial compliance" and 63% "no compliance." In contrast, the HEIS shows that over 6% is classified as "healthy", 86% as "requires changes" and 8% as "unhealthy." It is important to keep in mind though that these indices are not *per se* comparable because of their different foundations, but it is necessary to analyse them to understand which one fits better the purpose of the study, although both are basically intending to assess how closely people follow dietary recommendations. These differences can be further observed in Appendix C Figure C.8,

which are most likely stemmed by the lack of meat intake recommendations in the index of Chile.

The indices do not show many differences when analysed at a macro zone level, but when observed at a regional level (see Figure 3.1.3), for instance, the region XIV is the one with the highest proportion of people under the category "no compliance" for the index of Chile, but when observing the index of Spain, that same region has the lowest proportion of people under the category "unhealthy". Interesting distinctions can be observed across the different socioeconomic levels, where level 1 is the one with the largest percentage of people under the category "satisfactory compliance" and the smallest for "no compliance", thus apparently healthier than the other socioeconomic levels, according to the Chilean index. However, when observing the Spanish index, strong differences do not appear (see Appendix C Figure C.9). When it comes to age groups, not important differences are found. These insights already reveal that there are contrasts in food intake across sociodemographic variables, which is a step closer to answering the research question.

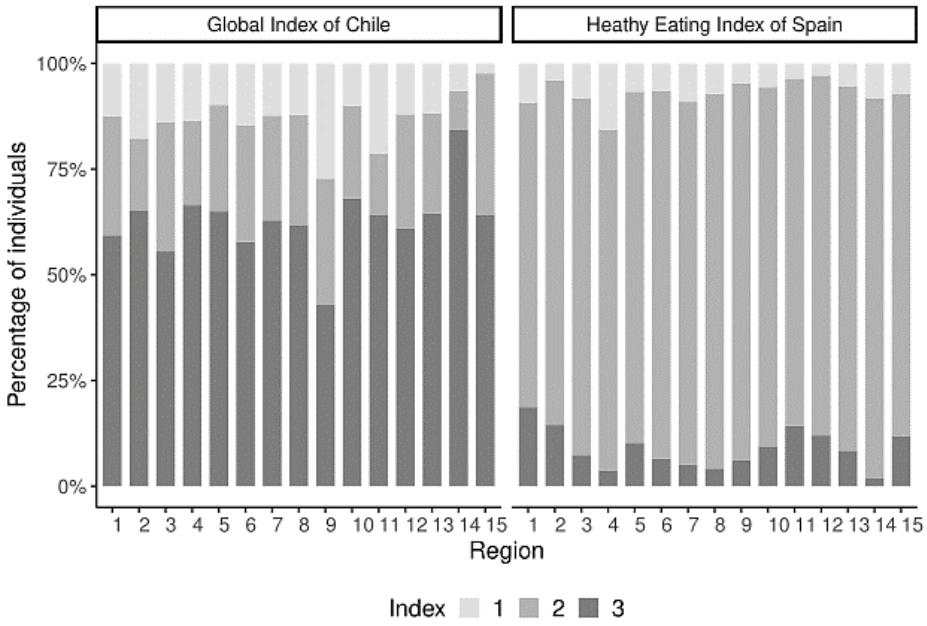

Figure 3.1.3: Comparison of healthy eating indices of Chile and Spain per Region. 1 corresponds to "compliance" or "healthy", 2 to "partial compliance" or "requires changes" and 3 to "no compliance" or "unhealthy."

3.2 Economic and Environmental Statistics per Region

A plethora of demographic, social, economic, and environmental indicators are available at a regional level. As exposed in Section 2.1, most studies focus on socio-demographic characteristics of the individuals or households but normally do not consider general attributes of the region to which the subject belongs to. However, because Chile is a country that has unique geography and is extremely diverse in terms of weather, natural resources and flora, and fauna, such differences across regions affect the way people live, where and how they work, and what they eat. Thus, in aims of finding possible relationships between regional features and diets, information is gathered to be analysed and later be included in the modelling process. The variables per Region potentially relevant for this thesis is shown below in Table 3.2.1.

Table 3.2.1: Variables representing regional features

Variable	Description
RURALITY	Percentage of rurality 2017. Rurality corresponds to areas with less than 1.000 inhabitants or between 1.001 and 2.000 inhabitants with less than 50% of them employed in secondary or tertiary activities. *Source*: prepared by the Office of Agricultural Studies and Policies (OASP) based on information from the XIX National Population Census and VIII Housing (2017) carried out by the National Institute of Statistics (NIS).
AGRI_PEOPLE	Percentage of people employed in agriculture, livestock, silviculture and fishing industries first trimester 2020. *Source*: NIS, Regional Employment Trimestral Series.
AGRI_SURFAFE	Land use for agricultural activities in hectares 2007. *Source*: prepared by the OASP based on information from the VII National Census of Agriculture and Forestry (2007) carried out by the OASP and NIS.
CO2_EMISSIONS	Total emissions of carbon dioxide equivalent (tCO2eq) in 2016. Carbon dioxide equivalent definition: "measure used to compare the emissions from various greenhouse gases based upon their global warming potential" (OECD, 2013). *Source*: National Greenhouse Gas Inventory System, 2018.
MAX_TEMP	Average maximum temperature in the hottest and coldest seasons (varying according to each Region) in Celsius degrees. *Source*: weatherspark.com, tool created by Cedar Lake Ventures[1] that compiles statistical analysis of historical weather reports and models from 1980 to 2016.

[1] About Weather Sparks, n.d. Retrieved May 10th, 2020, from https://weatherspark.com/about

The choice of features is driven by the intention of capturing the type of food production and lifestyle of the people that dwell in each region. For instance, regions with higher temperatures are more suitable for the production of fruits, whereas regions with lower temperatures have more rain and, therefore, more prairies that are appropriate for cattle. CO2 emissions may give a glimpse of the activity of the region and thus the level of development. When it comes to temperature, it is seen that northern regions have a higher maximum temperature compared to the southern regions, but the values of the rest of the features vary across regions (see Appendix D Table D.1).

Correlations were calculated between the variables (see Figure 3.2.1), which become relevant when implementing linear regressions and to understand the variable selection process performed by some algorithms. First of all, when it comes to the agriculture-related variables, RURALITY and AGRI_SURFACE present a correlation of 0.48, whereas RURALITY and AGRI_PEOPLE have a correlation of 0.78. Additionally, there are very low correlations between the temperature and such variables. Lastly, CO2_EMISSIONS are negatively correlated with RURALITY and AGRI_PEOPLE, -0.39 and -0.54 respectively, whereas with AGRI_SURFACE the correlation is almost zero and with MAX_TEMP is 0.56.

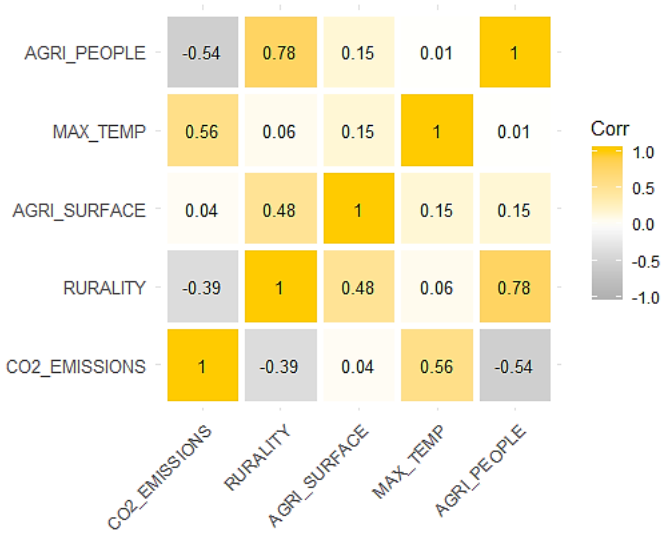

Figure 3.2.1: Correlation matrix for regional variables

3.3 Environmental Footprints

There is a wide variety of databases that show the environmental footprint of several foods. The environmental footprint can be found in terms of carbon footprint, water footprint, energy footprint, ecological footprint, land use, among many others. For simplicity purposes and due to the limited and non-comparable databases found, the environmental footprint of this study is limited to the carbon and water footprints. Carbon footprint is measured in grams of carbon dioxide equivalent (gCO2eq) per kilogram or litre of food, and water footprint is measured in litres of water resources used per kilogram or litre of food.

Because the consumption of products is mixed between imported and local foods, international and Chilean LCA databases are considered in this study. Notwithstanding, there is not much publicly available aggregated information comprising footprints and is quite atomized per country and industry, making it difficult to gather all the data, thus the focus is put on few sources of information. For instance, as exposed in Section 2.2, Food Ecobase is the most important source of LCA for Chile comprising 16 products, but one of the main drawbacks is the lack of variety of products and its high specificity, since, for example, it does not include beef meat or vegetables but it does include "dried apples" or "frozen breast chicken." In spite of that, important products are still present, such as milk, chicken, and pork, for which averages were calculated over its respective sub-categories. Furthermore, other specific products like Gouda cheese, salmon, mussels, and some fruits served as a reference for the impact of the mother category to which they belong to. The access to this calculator was obtained by contacting directly one of the professionals involved in the project.

Another source of information is the Barilla Center for Food and Nutrition. This institution has several publications and one of them includes carbon footprint, water footprint, and ecological footprint of 22 categories of food (BCFN Foundation, 2015). The major advantage of this dataset is that it comprises other relevant categories, like

"breakfast cereals", "butter", "sweets" and "cookies", besides the expected ones such as "milk", "beef", "poultry", "vegetables" and "fruits", and so on. A drawback is that there is not much detail on the methodology to calculate the footprints.

Lastly, Poore and Nemecek (2018) published a study where they consolidated data comprising five environmental indicators of thousands of sources worldwide. As a result, a database with the impact of 52 food items/categories was provided. To obtain this final and representative indicator, each observation was "weighted by the share of national production it represents, and each country by its share of global production." What is interesting from this database is that some predominant products have its own calculation of impact, excluding them of their main category, for the reason that they are highly consumed, namely apples, bananas, potatoes, tomatoes, and also similar fruits are grouped in sub-categories, such as "citrus fruits" and "berries and grapes." However, the water footprint is weighted according to the scarcity, thus from this source, only carbon footprint is being considered.

Additionally, other national sources of information were appealed to confirm the footprints of meat and dairy, foods with the biggest differences between international and national values. One of them is a document developed by the Institute of Agricultural Research of Chile (2013), where one of its chapters is dedicated to the water footprint of bovine meat and milk of the regions IX and X, from which the average between the two is used. Additionally, further detail for the water footprint of certain fruits, namely avocado, apples, peaches, prunes, and citrics is obtained from the same document. Another source of milk carbon and water footprints is the Sustainability Report by Manuka (2019), where the highest farms' values are used.

For most of the products, a simple average is calculated between the databases to obtain the reference footprint, but a special focus is put on the products that presented large differences when comparing global averages and Chilean footprints, this due to the reason

that production, transportation and consumption processes vary across countries. Thus, the ratio between the consumption of imported and local foods was necessary to calculate weighted footprints for certain categories of products. In general, regarding the consumption of all types of meats together, around 50% is local and the remainder local (Fernandez, 2017). When it comes to dairy, with special focus only on milk and cheese, the consumption of nationally-produced dairy is estimated based on the ratios between importation and the difference between local production and exportation (Office of Agricultural Studies and Policies of Chile, 2018). This results in 52% of milk being national and 48% imported, and for cheese only 4% is national, thus most of it is from international origins.

The final data of carbon and water footprints comprises 48 items, some of them corresponding to specific food items and other corresponding to categories (Appendix E Table E.1). This compilation did not cover all of the food groups that take part in the NFCS, thus some values were simply extended and used as a proxy for other food groups. To name a few examples, the footprint of the group "processed meats" from the NFCS is an average of the values for all the meats, the group "milk derivatives" from the NFCS takes the footprint of "liquid milk" and "ham" from the NFCS takes the value of "sugar."

To calculate the total footprint of a subject, the carbon and water footprints are multiplied by the consumption of the food group or item. Since the objective of this study is not to precisely quantify the footprints of the products but rather to create a model to identify the most relevant groups of people with diets that are not sustainable, an index was constructed based on the relative footprint concerning the rest of the individuals for both carbon and water independently. The individuals are classified according to the quintile they belong to, thus the index has five categories: 1, 2, 3, 4, and 5. For the Carbon footprint, it can be observed in Figure 3.3.1 that the average footprint increases exponentially as the levels increase, whereas for water it is more subtle.

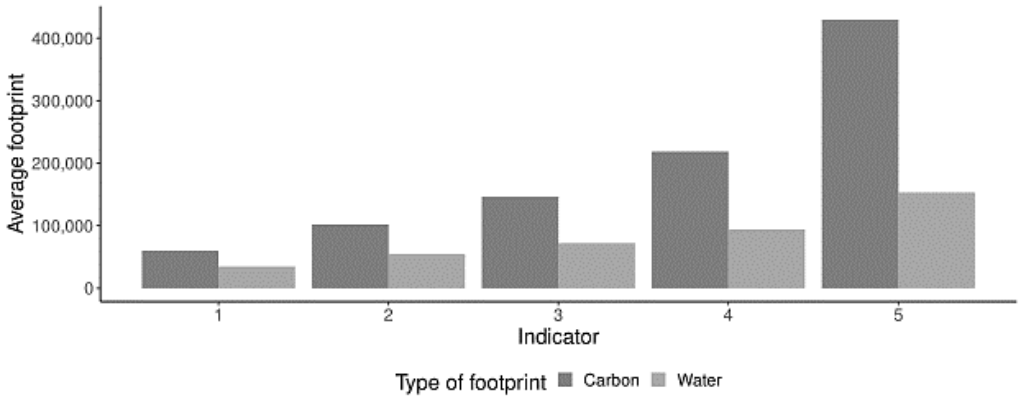

Figure 3.3.1: Carbon and Water Footprint indices

Nevertheless, interesting distinctions can be appreciated across macro zones; for instance, the South macro zone presents the smallest proportion of people classified in lower levels for both footprints, followed by the South Centre macro zone in the case of the carbon footprint and by the Metropolitan macro zone in the case of water footprint, whereas the North macro zone is the one that presents the smallest proportion of people classified under higher levels (Figure 3.3.2: Footprint indices across macro zones 3.3.2).

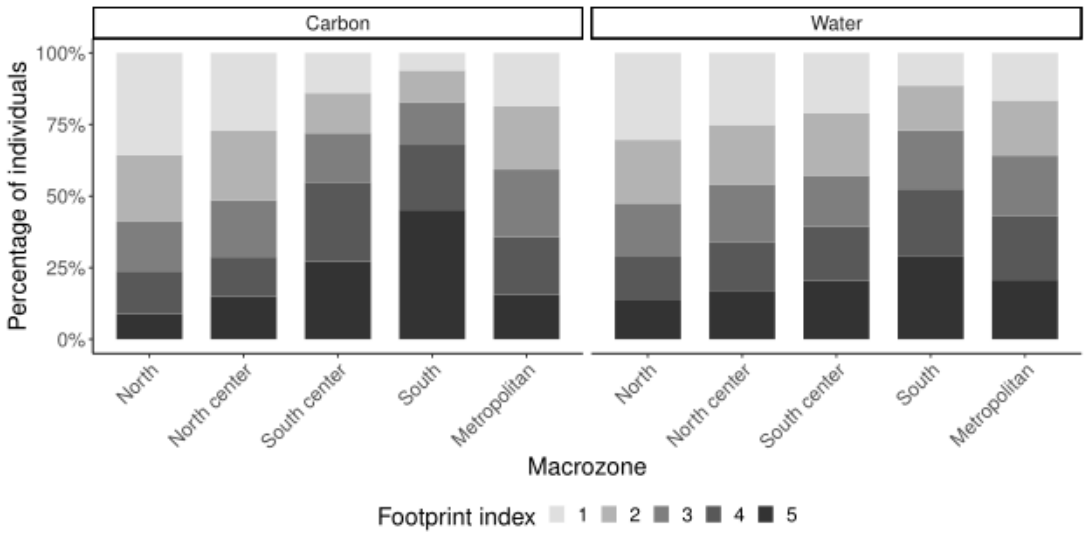

Figure 3.3.2: Footprint indices across macro zones

Regarding the relationship between age and footprint, there is a noticeable larger proportion of individuals classified in the category 1, both for water and carbon footprint,

for the older people and also slightly for the younger, contrasting with the middle-age groups that present a higher number of people that belongs to the categories 4 and 5 (see Appendix E Figure E.1). On the other hand, the higher the socioeconomic level, thus closer to 1, the smaller the proportion of people with larger footprints (see Appendix E Figure E.2).

Lastly, an interesting phenomenon is observed in the footprint indices when contrasted with both healthy eating indices. In the case of the GI of Chile, the lower compliance to the dietary recommendations, the lower the footprint or, in other words, more people classified as 1 or 2 for water and carbon footprints. However, the relationship is the opposite with the HEIS as illustrated in Figure 3.3.3. This may occur because the Chilean index does not consider meat intake when assessing the compliance, which happens to be among the food groups with the highest environmental impacts, thus it overlooks this aspect when in the case of the Spanish index this is taken into consideration. For this last reason, the HEIS is the index finally considered in this study.

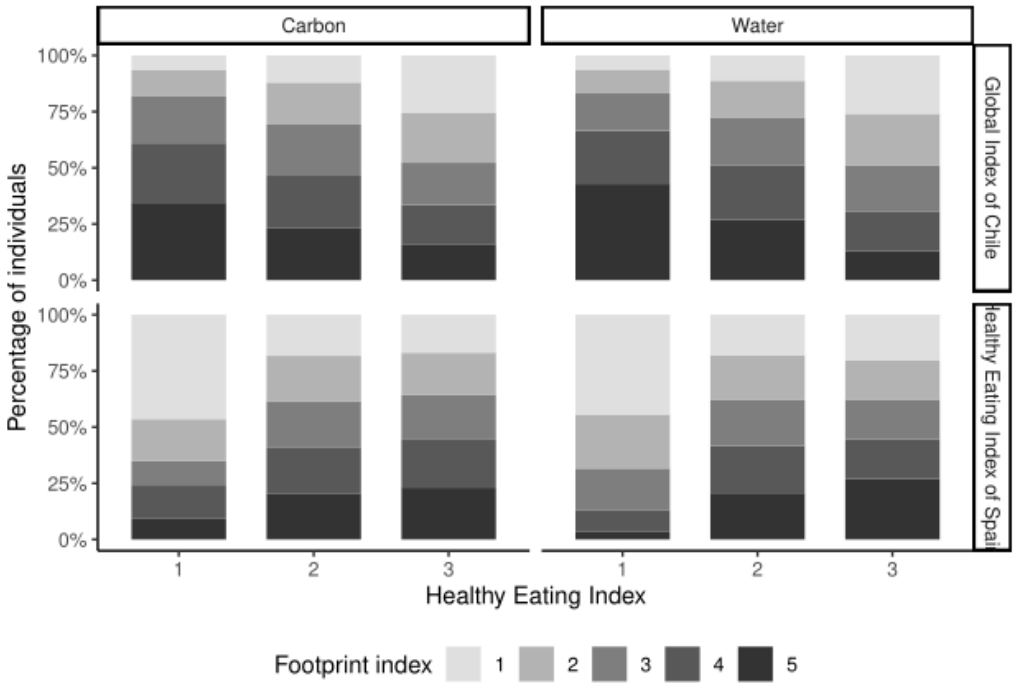

Figure 3.3.3: Healthy Eating index of Chile and Spain versus Footprint indices

4. Methodology

Two unsupervised learning methods are applied in order to find representative clusters with characteristics in common. Firstly, dimension reduction is applied by means of Principal Components Analysis (PCA) to subsequently apply the clustering method called k-Prototypes, used to construct the final groups. Afterward, relevant associations between the discovered clusters and independent variables, such as sociodemographic, are unveiled through Multinomial Logistic Regression (MLR) and Random Forest (RF).

4.1 Principal Components Analysis

PCA is a method that consists in the creation of a smaller number of new variables that in conjunction intend to represent the data in the best possible way, or in other words, these new variables summarize the original variables trying to account for most of the variability of the original information (James et al. 2017). These new variables are called principal components and each of them is a normalized linear combination of a set of variables $X_1, X_2, \dots X_p$. To better illustrate this, the linear combination of the first principal component is as follows:

$$Z_1 = \varphi_{11}X_1 + \varphi_{21}X_2 + \varphi_{p1}X_p \qquad (4.1.1)$$

This means that $\sum_{j=1}^{p} \varphi_{j1}^2 = 1$ and these elements $\varphi_{11}, \dots, \varphi_{p1}$ correspond to the component loadings and are the correlations between the original p variables and each component. To compute the principal components, the original features need to be normalized and then the loadings vectors of the first component Z_1 can be obtained by solving the optimization problem

$$P(\varphi_{11}, \dots, \varphi_{p1}) = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\varphi_{j1}x_{ij}\right)^2\right] \; subject\ to \; \sum_{j=1}^{p}\varphi_{j1}^2 = 1 \qquad (4.1.2)$$

where n is the total number of observations. The expression is equivalent to z_{i1}, thus the optimization function is simply maximizing the variance of the n values of z_{i1}, known as scores of the first component.

After finding the loading vector of the component Z_1 with the maximum variance, the second component Z_2 can be constructed. The second component is also a linear combination of the features $X_1, X_2, \ldots X_p$ that has the maximum variance but is uncorrelated to the component Z_1. The construction of the subsequent components Z_j follows the same logic of comprising the highest variances while being uncorrelated to the previous component Z_{j-1}.

To effectively reduce the dimensional space and not lose important information, a certain number of components need to be chosen. The most common method is by checking the Variance Accounted For (VAF) of each component in a scree plot. The VAF is the percentage of variance explained by each component, hence the cumulative VAF of all the components adds up to 1. In the scree plot can be observed an elbow where after certain components not much VAF is added. In addition to this, components with eigenvalues >1 is a complementary criterion to select them. Overall, the idea is to choose a small number of components that capture a significant percentage of the variance of the original data set. Finally, the summarized data will comprise the chosen principal components, thus new variables and the values of each observation for each new variable will correspond to the principal component scores.

4.2 k-Prototypes

There are several methods to cluster data into homogeneous groups in the sense that the objects within them are similar and the clusters are dissimilar among each other. However, some methods are not adequate for mixed-types of data and cannot handle large datasets well. For instance, the Hierarchical Clustering method can handle both categorical and numerical data when the distance between the observations is calculated using Gower's dissimilarity measure (Gower, 1971), but the clustering procedure can be computationally intensive and other methods, namely k-Means, are more appropriate for these tasks.

k-Prototypes is an extension of the k-Means clustering algorithm by combining it with the k-Modes algorithm, because one of the drawbacks of both of these algorithms is that the

first one works only with numerical data, by using the mean of the clusters as centroids, and the second one works only with categorical data, by using the mode as centroids of the clusters and updating them with a frequency-based method. Thus, k-Prototypes utilizes a combined dissimilarity measure to cluster observations with both types of data, numerical and categorical (Huang, 1998), where the centroids are here known as cluster prototypes, hence is a representative-based algorithm.

The objective function of k-Prototypes aims to minimize the distances or dissimilarity between the k cluster prototypes and the observations within the cluster k^{th}, given by:

$$E = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij} d\,(x_i, \mu_j)\ , \qquad (4.2.1.)$$

where u_{ij} is the $n \times k$ binary matrix that represents whether the observation x_i, with $i = 1, \dots, n$, belongs to a cluster or not, thus satisfies $\sum_{j=1}^{k} u_{ij} = 1,\ \forall i$, and μ_j is the cluster prototype, with $j = 1, \dots, k$. $d(\,)$ is the distance function corresponding to a weighted sum of the distances for numerical variables and categorical variables, defined by:

$$d(x_i, \mu_j) = \sum_{m=1}^{q} (x_i^m + \mu_j^m)^2 + \ \lambda \sum_{m=q+1}^{p} \delta(\,x_i^m, \mu_j^m)\ , \qquad (4.2.2)$$

where the variables with m between 1 and q are numerical, and the remainder with m between $q+1$ and p are categorical. The expression of the distance for the numerical variables corresponds simply to the squared Euclidean distance and the function $\delta(\,)$ in the expression for the categorical variables counts the mismatches, so it is based on the mode, hence $\delta(a, b) = 0$ for $a = b$, and $\delta(a, b) = 1$ for $a \neq b$ (Szepannek, 2018). λ controls the trade-off between the importance of categorical and numerical data, so when λ is larger, numeric attributes are less relevant, so observations of two different classes are easier to separate based on the categorical variables, which can lead to higher accuracy levels. In general, the value of lambda provides better results when guided by the average standard deviation σ of the numeric features or $\frac{1}{3}\sigma \leq \lambda \leq \frac{2}{3}\sigma$ (Huang, 1997). However, another alternative that provides a good balance between the two types of variables can be found by correcting the standard deviation of the numerical variables σ by the average concentration of the categorical variables, given by $\lambda = \frac{\sigma}{h_c}$ where the concentration h_c is

represented by the average of $h_m = 1 - max_c p_{mc}$ $\forall$ $m = q + 1, \ldots, p$, thus the average of the maximum proportions given by the class c for each categorical variable (Szepannek, 2018). The choice of using the largest concentration per categorical variable rather than averaging all concentrations is because it is simpler, effective and makes sense when the categories have only a few classes. Hence, if the categorical variables are strongly inclined towards one of the classes, lambda becomes larger, meaning that categorical variables, in such cases, provide more relevant insights when assigning observations to a cluster. On the other hand, when numerical variables present a high variability, lambda also becomes larger thus attributing more importance to categorical variables due to the instability of the numeric features.

When performing k-Prototypes algorithm, another parameter that needs to be provided is the number of clusters to be formed. A low k may provide a good generalization of the data but fuzzier groups, whereas a large k could provide more distinct groups but make the interpretation harder. The goal is to find a number of k clusters that results in a low total within dissimilarity (TWD), corresponding to the objective function, and that also bestow interesting and interpretable groups. Nonetheless, logic and math suggest that the higher the number of clusters, the lower the total within dissimilarity, thus it is necessary to appeal to another tool. Similarly, as for PCA, scree plots prove to be useful when searching for a negligible decrease in the TWD after a certain k. Additionally, the silhouette index can be more precise to find the optimal number of clusters. The silhouette index is given by:

$$Silhouette = \frac{1}{n}\int_{i=1}^{n} \frac{b(i)-a(i)}{max(a(i),b(i))} \qquad (4.2.2)$$

where $a(i)$ is the average distance of the i^{th} instance to the other instances and $b(i) = \min(d(i, C))$, thus the distance to the nearest cluster excluding the own cluster. The maximum value of the index indicates the best number of clusters (Thinsungnoena et al., 2015).

The algorithm iterates re-computing the cluster prototypes and assigning the clusters until the objective function is met. Nonetheless, there is the possibility of obtaining a local minimum, thus it is recommended to do a number of random initializations and choose the lowest result among them.

4.3 Multinomial Logistic Regression

MLR is a generalized linear regression for classification purposes. Working as an extension of the Logistic Regression, MLR is based on the estimation of the maximum likelihood to predict the probabilities of all m classes of a categorical dependent variable Y with more than two possible outcomes.

The MLR estimates $m-1$ logit equations, thus it is a probabilistic model, using as a reference one of the categories of the dependent variable. It assumes that the log-odds of each category follow a linear model, thus

$$\log(\Pr(Y_{ik})) = \beta_{0k} + X_i\beta'_k \qquad (4.3.1)$$

where $\Pr(Y_{ik})$ corresponds to the probability of the k^{th} category for the i^{th} observation and β_k is the vector of coefficients estimates for the vector of predictor variables X_i. Therefore, the probabilities of the outcomes can be expressed as

$$\Pr(Y_{ik}) = \Pr(Y_i = k|x_i; \beta_1, \beta_2, \dots, \beta_m) = \frac{\exp(\beta_{0k}+X_i\beta'_k)}{\sum_{j=1}^{m}\exp(\beta_{0k}+X_i\beta'_k)} \text{ with } k = 1, 2, \dots, m \qquad (4.3.2)$$

The most relevant output of the MLR is the coefficients because they capture the relationships between predictor variables and the dependent variable. Furthermore, the significance of the predictors for each output category indicates which features are the most relevant.

Some considerations need to be taken into account when training a linear model. Firstly, to make the coefficients comparable and more interpretable, numeric variables should be standardized before training the model. Secondly, linear regression does not handle well-

correlated features, thus this needs to be analysed beforehand to choose a curse of action, being either not including strongly correlated variables or by performing variable selection with the penalization of the model, for instance.

To check the predicting capability of the model and to confirm that it is not overfitting the data, a training set should be used to train the model, which usually comprises around 75% of the observations. Subsequently, the remainder observations, corresponding to the test set, can be used generate predictions with the model and check the accuracy, which corresponds to the percentage of correctly classified observations.

4.4 Random Forest

RF is a type of machine learning Ensemble method (James, G., Witten, D., Hastie, T., Tibshirani, R., 2017) that combines Decision Trees (DT) to improve the predicting capability of the model, thus it is a supervised learning method. It can serve for two main purposes; first, to predict or classify observations and second, to understand the associations between the dependent variable and the independent or predictor variables.

To understand RF, it is first necessary to understand how DT work. DT is a procedure based on recursive binary splitting, aiming to classify observations or predict values. Their structure is similar to a tree, hence its name, because they have splitting points, called nodes, and branches. The trees start with a single root node, corresponding to the non-split data, and end in terminal nodes, corresponding to the split data. The set of predictor variables is considered at each node to find the one and its corresponding cut point that best split the space with respect to a given measure of "split quality", such as the Gini Index in the case of categorical dependent variables. This is recursively repeated until the quality measure cannot improve further.

DT alone can result in an inaccurate model with high variance, tending to overfit the data if not pruned appropriately. Besides, they suffer from selection bias towards the predictor variables that offer many possible splits, namely numerical or categorical features,

affecting the interpretation of the model. However, the variance can be reduced by using an Ensemble method such as RF, which is the method applied in this thesis.

Employing Bootstrapping Aggregation (Bagging), RF draws several samples with replacement with the same number of observations n from the original training to construct each tree. In complement to this, only a subset of the p predictors is deemed to perform the splits, this not only helps to better handle correlated variables, but it also provides a solution to the issue of selection bias towards the best features, otherwise, the resulting trees could be too similar to each other.

Because in this study RF is used to classify observations, Gini Index is the measure for split quality chosen. The Gini Index is calculated at each node, given by

$$GI = 1 - \sum_j p_j^2 \ , \qquad (4.4.1)$$

where p_j corresponds to the probability of occurrence of the class j. Gini index takes on a small value if all of the p_js are close to zero or one or, in other words, if the node contains predominantly observations from a single class, hence it is purer.

The construction of the model firstly starts by splitting the dataset into a Training set and Test set. It is recommended for the Training set to comprise around 75% of the observations. The following step is to draw the "bootstrapped" samples from the Training set to build the trees. With this, the first parameter that is optimized is the number of trees that are going to be constructed, which is equivalent to the number of bootstrapped samples with replacement needed. According to the probabilities resulting from the "bootstrapped" samples, approximately $n/3$ observations do not take part in any of the training samples, and these are named out-of-bag (OOB) observations. These are used to calculate the OOB GI.

The algorithm for growing each tree stops when the OOB GI no longer improves. The number of trees B that converge to the best index will correspond to the optimal, nonetheless, due to the computational heaviness of the algorithm, the number of trees can be set to a number such that after it the error or index does not improve considerably.

The most commonly occurring class among all B or majority vote will correspond to the overall prediction for a given observation.

The number of predictor variables that are going to be considered for each split corresponds to the second parameter to be optimized. Usually corresponds to $\sqrt{m}$ or $m/3$, however, it is tuned via grid-search, looking for the one that yields the best OOB GI. After finding the tuned parameters and the model is ready, the Test set is used to check for the index stability with 10-folds cross-validation.

The objective of utilizing RF to predict classes, in this case, is to understand the relationship between the predictor variables and the dependent variable by using a robust model. However, in contrast to regression models, RF does not provide a direct interpretation of the impact of the variables over the classification task, due to the complexity of the operations that occur behind the model, being considered a Black Box model. A way to calculate the variable importance is by assessing how much the GI increases when the values or classes of the predictor variable in study are shuffled, the rest remaining the same, thus taking a permutation-based approach. The more the GI increases, the more relevant is the predictor. It is pertinent to keep in mind that correlation between features is detrimental when assessing variable importance, due to the fact that even when the values of a specific predictor are permuted, there still will be one or more features that capture or explain the effect over the dependent variable, at least to some extent, thus resulting in smaller importance.

5. Results

Several steps were performed in order to get closer to answer the research question, depicted in the flowchart in Figure 5.0.1. The first task before clustering the observations with the k-Prototypes algorithm is to perform dimension reduction of the food items through Principal Components Analysis in aims to find relevant dietary patterns. Afterward, clustering is performed to classify individuals into based on the dietary patterns, environmental footprint indices, and a healthy eating index. Lastly, supervised

methods are applied to discover associations between clusters and sociodemographic and regional variables.

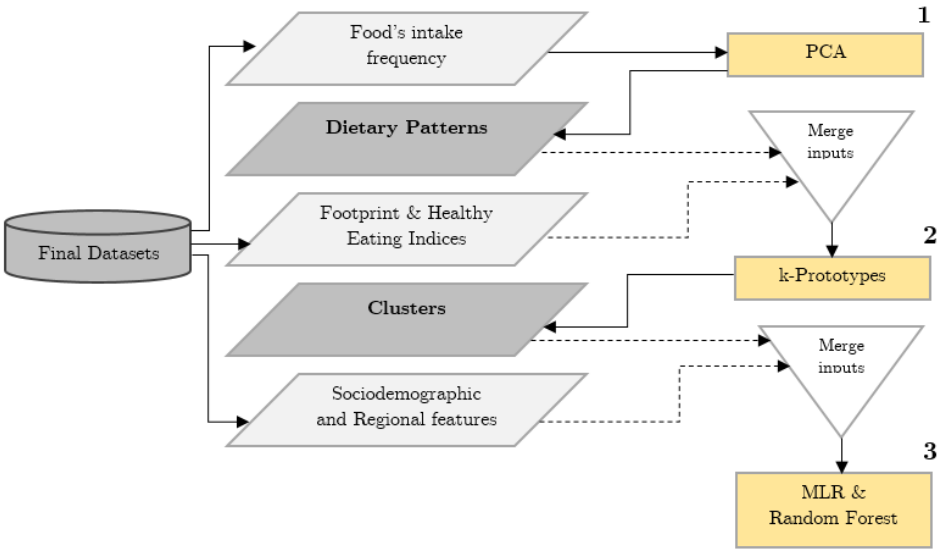

Figure 5.0.1: Flowchart of the methods applied to the datasets

5.1 Dietary Patterns

Principal Components Analysis is performed over the individuals' intake frequency of the foods aggregated at level 2, thus 68 items, resulting in five principal components or dietary patterns. These five components have a cumulative proportion of variance explained of 19.2% in total, thus between common ranges for this kind of data as mentioned in Section 2.1. The selection is based on the elbow in the scree plot (see Appendix F Figure F.1), eigenvalues greater than one, and interpretability. Overall, from the sixth component onwards, the percentage of cumulative variance does not increase considerably and the food composition does not vary in a way that allows identifying those components as distinctive dietary patterns, whereas the first five principal components are by themselves representative diets in the sense that they are relevant, they differ from each other, and some of them can stem from intuition. However, it is important to keep in mind that each observation will present a combination of the five components that will determine a proxy to their real diet pattern. In other words, now each observation has a score for each principal component, which are the new variables that summarize the original

information. The description of each dietary pattern is as follows, according to the loadings of the principal components (see Appendix F Table F.1).

The first principal component, named "**Diverse and light foods**," is characterized by higher consumptions of traditionally considered healthy products such as poultry, light yogurt, light jam, whole wheat pasta, skim milk, integral rice, white meat, nuts, light soft drinks, sweeteners, fruits and vegetables in general and olive oil. However, this dietary pattern still presents moderate consumption of most of the remaining food groups, namely cheese, and some of them not necessarily healthy, such as the other animal meats, including processed meats. Hence the name "Diverse and light inclined" for this component.

The second component is named "**Low intake and light foods**" due to its very low consumption of all types of meats but also of fruits and vegetables. Similarly to the "Diverse and light" pattern, its frequency intake of bread, yogurt, pasta, bread, and cookies is higher for their light or whole versions, rather than for the regular.

Higher consumption of lamb, pork, other meats, animal butter, eggs, fish, dried and fresh fruits, vegetables in general, other grains, sweet pastries, *manjar*, legumes, sugar, potatoes, white rice, wine, and tea, can be observed in the third component, called "**Southern traditional and animal products intensive**." This pattern resembles a traditional diet from the southern area of Chile, thus its name.

The fourth component, named "**Meat and alcohol inclined**," also shows a diverse diet in the sense that loadings are moderate in most of the foods, however, it can be highlighted that there is a stronger intake of processed meats, beef, pork, eggs, beer, liquors, wine, sunflower oil, white rice, regular pasta, cream, cheese, and sauces and dressings. On the other hand, rather low consumption is presented for vegetables and fruits.

The last component is named "**Sugar intensive**" and it presents higher loadings for chocolate, sweet pastries, regular soft drinks, sugar, *manjar*, ice cream, regular jam, regular yogurts and dairy desserts, breakfast cereals, and fruits. Additionally, it can be observed that there is low consumption of all types of meat.

5.2 Clusters

By means of k-Prototypes applied over the five dietary patterns, carbon and water footprint indices, and the Healthy Eating Index of Spain (HEIS), thus eight variables, clusters are formed. In this case and according to the Silhouette index, the best option is six clusters. As a summary, clusters are characterized by a description according to the basis variables used to construct them and also by their size and sociodemographic features. The footprint indices of each cluster can be observed in Figure 5.2.1. Due to the nature of the data, most of the people belong to the class "requires changes" when it comes to the HEIS (80% to 95%) for all clusters and are from the Metropolitan region (80% to 95%), equivalent to the Metropolitan macro zone or region XIII, so only interesting distinctions are mentioned in the following descriptions.

Cluster 1 comprises 963 individuals, corresponding to 20.7% of the sample, whose diet is principally "Meat and alcohol inclined," however it does present some characteristics of the other diets "Low intake and light foods" and "Southern traditional and animal products intensive," thus quite mixed. The consumed food items can be observed in more detail in Appendix G, where the frequency consumption is more or less around the average of the whole sample, especially for different types of meats and carbohydrates-packed foods, such as rice, potatoes, and pasta, however, it can be highlighted that there is slightly lower consumption of fruits and vegetables in general. Its carbon and water footprint are mostly concentrated around the categories 2, 3, and 1, where around 40% of the individuals belong solely to category 2, both for water and carbon, thus it could be said that their footprint is not that high. Its healthy eating index is around the average with 85% of people under the category "requires changes." The average age is around 39 years old and their Body Mass Index (BMI) is around the general mean. It is the second cluster with the highest percentage of rurality, on average, with 14.7%. Regarding their geographic distribution, the individuals from this cluster are more or less spread across macro zones, still concentrating their majority in the Metropolitan macro zone with 34.9%. Their socioeconomic level is medium-low, with 37.1% of people that belong to level 4, and around 21% that belong to levels 2 and 3.

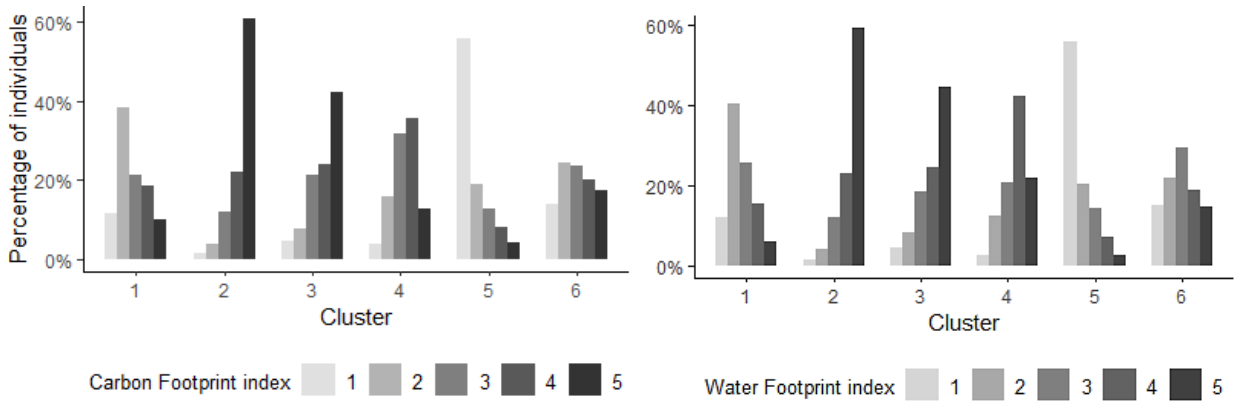

Figure 5.2.1: Water and Carbon Footprint Indices across clusters

Cluster 2 includes 547 observations, equivalent to 11.7% of the sample. Compared to cluster 1, its main dietary pattern is also "Meat and alcohol inclined," but to a larger extent. Furthermore, this cluster's diet is not related to the "Low intake and light foods" and "Southern traditional and animal products intensive" diets, and even less to the "Sugar intensive." More specifically, individuals from this cluster present a higher intake of liquors, beer and wine, white rice, regular soft drinks, coffee, different types of meat (except for lamb), and also carbohydrates-packed foods, when compared to the average or other clusters. As intuition suggests, this cluster is not the most sustainable or healthy one. Around 60% of the people fall under category 5 when it comes to both carbon and water footprints. Furthermore, it has the highest proportion of people classified as "unhealthy," with 18.5%. This group is slightly young, with an average age of almost 35 years old and an average BMI of 27.5. This is the cluster with the second-highest proportion of people that belong to the South macro zone, 23%, and is the one that breaks the balance between men and women, comprising 68% of males. They have a mid-socioeconomic level, with around 23% to 30% of people belonging to each of the levels 2, 3, and 4.

Cluster 3 is composed of 591, thus 12.7% of the sample and its main dietary pattern is the "Southern traditional and animal products inclined," followed by "Diverse and light foods" and "Meat and alcohol inclined." At the same time, this cluster is the furthest from the "Low intake and light foods" pattern. Foods such as potatoes, cheese, alcohol, honey,

jam, butter and margarine, legumes, whole milk, eggs, meats, fruits, and vegetables in general, wheat, cereals, and nuts, appear to have larger frequency intakes when compared to the average and other clusters. Their footprint levels are rather high, with more than 42% falling under category 5 and 23% under category 4 for both carbon and water. The people from this cluster are from regions that on average have the highest percentage of rurality, 16.8%, and 29% are from the South macro zone, this being the highest proportion among all clusters. Their age is above the average, with a mean of 47.3 years old. Their BMI is around the average, 27.2, and almost 95% of people are classified as "requires changes." Their socioeconomic level is rather low, with 12.9% of people in level 5, 39.4 in level 4, and 21.5 in level 3.

For **Cluster 4**, which comprises 731 observations, their main dietary pattern is "Sugar intensive" and it is rather far from the "Low intake and light foods" diet. Their consumption of fruits and vegetables is close to the average, but in the case of processed meats, beef, poultry, pork, chocolate, regular soft drinks, ice cream, whole milk, sweet pastries, and regular yogurt, a higher frequency of consumption is observed. When it comes to carbon footprint, 31.7 of the observations fall under category 3 and 35.8% under category 4. In the case of water footprint, 20.7 present a level 3 and 41.4% a level 4. According to the healthy eating index, 7.8% are classified as unhealthy, thus slightly larger than the average, but, interestingly enough, they have the lowest average BMI, 23.7. However, this may go in line with the fact that this is the youngest cluster, with an average age of 23 years old. This cluster presents a high proportion of people that belong to the Metropolitan macro zone, 47.2%, followed by North Centre and South Centre macro zones with 18.7 and 16.7% respectively. Their socioeconomic level is relatively low, with almost 70% of people under the levels 3 and 4, but almost no people in level 5. This cluster stands out for presenting the largest amount of family members, 4.4 people on average.

Cluster 5 presents a diet mostly identified with the "Low intake and light foods" pattern, followed by "Southern traditional and animal products intensive," and 1,192 individuals. This cluster is characterized by presenting a low consumption frequency of sugar-packed

foods, but in general, they are slightly below the average for most of the food items, particularly for most types of meats, for which the intake is far less than one time per week. In some items, higher consumption is observed though, for example for other edible oils, for tea and the groups powder soups, desserts, and so on. It has the lowest environmental impact, with around 56% of people that fall under category 1 and around 20% under category 2, for both carbon and water footprints. According to the healthy eating index, this cluster has the lowest percentage of people under the category "requires changes", and the remaining observations are evenly split between "healthy" and "unhealthy." It has the highest proportions of people that belong to North and North Centre macro zones when compared to other clusters. This cluster has the highest average age, 49.3 years old, and one of the highest BMI, above 28. Almost 65% belong to the socioeconomic levels 4 and 5 together.

Cluster 6 comprises 639 observations and its main dietary pattern is the "Diverse and light foods" and also, but in a lesser way, the "Low intake, and light foods." Individuals from this cluster present, on average, higher consumption of olive oil, mineral water, integral rice, light soft drinks, fruits and vegetables in general, nuts, light cookies, skim milk, whole wheat bread and pasta, vitamins, and supplements and light yogurt. They have lower consumption of lamb and pork, sugar, regular soft drinks, regular pasta, potatoes, white bread, sweet pastries, whole milk, among others. This is the second cluster with the lowest environmental footprints levels, with 24.6% of people in level 2, 24.8% in level 3 and 13.9% in level 1 for carbon, and 29.3% of people in level 3, 21.8% in level 2 and 15.2% in level 1. The proportion of individuals classified as "healthy" is 16.6%, the highest among all clusters, even though their average BMI is rather high, also above 28, with an average age of 48.9. Half of the people in this cluster are from the Metropolitan macro zone, the metropolitan region, and more than 52% of people have a socioeconomic level of 1 or 2, this being the largest proportion when compared to the other clusters. Additionally, they have the smallest number of family members, 3.3 on average, and 78.9% of the people are women.

It is worth analysing how this classification of individuals would reflect on the population. By using the Expansion Factor per observation it is possible to estimate how many individuals of the population would belong to each cluster. From the proportions shown in Table 5.2.1, it can be derived that Cluster 2, which has high environmental footprints, is actually larger when observing the proportions in the population. Oppositely, Cluster 5, the most "environmentally-friendly" of the clusters, becomes less relevant in terms of size when comparing the population with the sample. The rest of the clusters present smaller differences when contrasting sample versus population, nonetheless, the general insight is that less sustainable and less healthy clusters are more relevant than what the sample suggests.

Table 5.2.1: Cluster's size in the sample and population

	Clust. 1	Clust. 2	Clust. 3	Clust. 4	Clust. 5	Clust. 6	Total
Sample	963	547	591	731	1,192	630	4,663
	20.61%	11.73%	12.67%	15.68%	25.56%	13.51%	100%
Population	3,279,409	2,455,110	2,009,596	2,621,935	3,528,955	1,763,602	15,658,607
	20.04%	15.68%	12.83%	16.74%	22.54%	11.26%	100%

5.3 Associations Between Clusters and Variables

In order to discover and understand significant relationships between the classification of an individual under a cluster and their respective sociodemographic features, first, two Multinomial Logistic Regression (MLR) are performed. The first one is to understand how the location of the individual in conjunction with personal characteristics affect the prediction of clusters. The second one is to capture the effect of the characteristics of the regions over the prediction of clusters. Afterward, Random Forest (RF) algorithm is computed to further confirm associations between the dependent variable and predictors and to explore the effect of the Expansion Factor.

5.3.1 Multinomial Logistic Regression

Before training the model, the sample is split into a train data set and test data set, comprising 75% and 25% of the observations respectively. Two MLR models are constructed. The first MLR trained is the one comprising the variables REGION, AREA, AGE, GENDER, SOCEC_LEVEL and TOT_MEM, called Model A. The train overall accuracy is 39.29% and then the model is used to predict over the test data set, resulting in an overall test accuracy of 37.99%, thus considerably better than chance. In Table 5.3.1.1 the prediction quality can be observed in more detail. The main insight that can be derived so far is that the independent variables are not able to predict satisfyingly observations that belong to Cluster 3. In spite of that, relevant relationships were unraveled.

Table 5.3.1.1: Confusion Matrix of the Test predictions Model A

		Reference					
		1	2	3	4	5	6
Prediction	1	44	26	25	22	24	14
	2	24	30	22	17	14	9
	3	7	1	14	2	4	0
	4	43	17	21	105	42	19
	5	89	34	68	31	188	74
	6	25	7	17	9	16	62
Accuracy		18.97%	26.09%	8.38%	56.45%	65.28%	34.83%

The resulting coefficients in Table 5.3.1.2 are obtained with Cluster 1 as reference class and the analysis is focused on the significant features. When it comes to relevant variables, it is worth mentioning the effect of the socioeconomic level, which shows that the lower the level, the lesser the odds of Cluster 6 being predicted when compared to Cluster 1, starting with a decrease of the log the odds of 1.10 and up to 3.02. The opposite effect is observed for Cluster 5, where socioeconomic levels 4 and 5 increase the log of the odds in 0.13 to 0.22.

Table 5.3.1.2: Multinomial Logistic Regression A results on the prediction of Cluster with sub-set of independent variables

	Clust. 2	Clust. 3	Clust. 4	Clust. 5	Clust.6
(Intercept)	-0.48 (0.71)	-1.20 (1.14)	-0.46 (0.75)	-1.76 (0.89)	0.64 (0.71)
REGION II	0.22 (0.78)	1.99 (1.21)*	1.00 (0.80)	2.58 (0.93)***	1.17 (0.79)
REGION III	0.27 (0.77)	0.89 (1.26)	0.51 (0.80)	2.08 (0.92)**	0.95 (0.78)
REGION IV	-1.13 (0.74)	1.29 (1.14)	-1.46 (0.79)*	1.55 (0.87)*	-0.07 (0.71)
REGION V	-0.73 (0.66)	0.97 (1.12)	-0.35 (0.71)	1.48 (0.85)**	-0.07 (0.67)
REGION VI	-0.73 (0.76)	1.92 (1.16)*	0.57 (0.75)	2.37 (0.88)***	0.70 (0.72)
REGION VII	-1.53 (0.70)**	1.13 (1.13)	-0.81 (0.72)	0.77 (0.86)	-0.51 (0.70)
REGION VIII	-1.03 (0.66)	1.63 (1.12)	-0.56 (0.70)	0.97 (0.85)	-0.10 (0.66)
REGION IX	-0.39 (0.73)	1.92 (1.15)*	-0.35 (0.77)	1.42 (0.89)	0.03 (0.75)
REGION X	0.28 (0.66)	1.41 (1.13)	-0.69 (0.72)	0.89 (0.86)	-0.09 (0.70)
REGION XI	0.70 (1.11)	2.73 (1.38)**	-4.78 (9.43)	1.05 (1.26)	-4.04 (9.31)
REGION XII	0.89 (0.70)	1.30 (1.17)	-0.21 (0.78)	0.64 (0.92)	0.52 (0.72)
REGION XIII	-0.34 (0.64)	1.32 (1.11)	0.03 (0.69)	1.54 (0.84)*	0.45 (0.64)
REGION XIV	-1.02 (0.74)	1.21 (1.15)	-1.37 (0.80)*	0.45 (0.90)	-0.50 (0.76)
REGION XV	-0.93 (0.95)	0.20 (1.54)	-0.34 (0.89)	2.15 (0.97)**	-0.61 (1.06)
AREA 1	0.01 (0.23)	-0.59 (0.20)***	-0.28 (0.20)	-0.48 (0.17)***	0.24 (0.27)
GENDER 1	1.24 (0.13)***	0.17 (0.13)	-0.02 (0.12)	-0.31 (0.11)***	-0.83 (0.14)***
AGE	-0.12 (0.07)*	0.43 (0.07)***	-1.00 (0.08)***	0.43 (0.06)***	0.50 (0.08)***
SOCEC_LEVEL 2	-0.24 (0.25)	-0.33 (0.26)	-0.15 (0.26)	0.44 (0.29)	-1.10 (0.21)***
SOCEC_LEVEL 3	-0.22 (0.25)	-0.12 (0.25)	0.36 (0.25)	1.22 (0.28)***	-1.30 (0.22)***
SOCEC_LEVEL 4	-0.48 (0.23)**	-0.42 (0.24)*	0.30 (0.24)	1.13 (0.27)***	-1.95 (0.21)***
SOCEC_LEVEL 5	-0.46 (0.29)	-0.55 (0.28)**	0.09 (0.30)	1.53 (0.30)***	-3.02 (0.35)***
TOT_MEM	0.19 (0.07)***	0.09 (0.07)	0.09 (0.07)	0.08 (0.06)	-0.17 (0.07)**

Note:
Logistic regression coefficients with standard errors in parentheses with Cluster 1 as reference category for the dependent variable; *p<0.2, **p < 0.05, ***p < 0.01

Regarding age, this variable is negatively related to Cluster 2, -0.12, and Cluster 4, -1.00. When it comes to the influence of the number of members of the household, higher numbers increase the probabilities of Cluster 2 being predicted, whereas for Cluster 6 are diminished. On the other hand, for gender, when the individual is male, the log-likelihood for Cluster 2 increases 1.24, whereas for Cluster 4 and Cluster 5 it decreases in 0.31 and 0.83 respectively.

Interesting effects are observed in geographic variables. For instance, for individuals that live in urban areas, the log probability for Cluster 3 is lessened in 0.59 and for Cluster 5 in 0.48. When it comes to specific regions, it is worth mentioning that regions from North

and North Centre present a positive association with Cluster 5, considerably increasing its probabilities of being predicted, when compared to Cluster 1. The effect of the Regions VI, IX, and XI over Cluster 2 is positive. Lastly, negative relations are observed between region VII and Cluster 2, decreasing its log of the odds in 1.53, and between region IV and Cluster 4, with a coefficient of -1.46.

The second regression, namely Model B, includes the variables AREA, AGE, GENDER, SOCEC_LEVEL, TOT_MEM, RURALITY, CO2_EMISSIONS, and MAX_TEMP. It is relevant to mention that due to the correlation issues, AGRI_SURFACE and AGRI_PEOPLE were excluded to not interfere with the interpretation of the results. The train accuracy resulted in 37.78% and the test accuracy in 36.88%. The confusion matrix is similar to the one from Model A and the lax predictive capability towards Cluster 3 increases (see Table 5.3.1.3). The relationships between the dependent variable and AREA, AGE, GENDER, SOCEC_LEVEL, and TOT_MEM continues to be mostly the same, but one important difference is that now most of the intercepts are significant, thus they are now capturing information that was contained in the excluded variables, as shown in Table 5.3.1.4. This time, attention is put over the new variables that intend to capture the characteristics of the regions. For instance, regions with higher temperatures are positively associated with Cluster 4, with a coefficient of 0.18. When it comes to CO2 emissions, regions with larger values decrease the probability for Cluster 2 being predicted, but the opposite occurs for Cluster 4 and Cluster 5. Regarding RURALITY, a negative relationship is found with Cluster 2, decreasing the log of the odds in 0.40, and increasing them for Cluster 3 in 0.15, having as reference Cluster 1.

Table 5.3.1.3: Confusion Matrix of the Test predictions Model B

		Reference					
		1	2	3	4	5	6
Prediction	1	35	20	16	15	17	19
	2	27	31	16	15	11	8
	3	3	0	3	0	3	1
	4	43	20	26	109	46	16
	5	99	36	88	39	191	73
	6	25	8	18	8	20	61
Accuracy		15.09%	26.96%	1.79%	58.60%	66.32%	34.27%

Table 5.3.1.4: Multinomial Logistic Regression B results on the prediction of Cluster with sub-set of independent variables

	Clust. 2	Clust. 3	Clust. 4	Clust. 5	Clust.6
(Intercept)	-1.09 (0.31)***	0.26 (0.28)	-0.73 (0.30)***	-0.38 (0.3)	0.77 (0.31)***
AREA 1	0.17 (0.22)	-0.64 (0.19)***	-0.22 (0.20)	-0.45 (0.16)***	0.33 (0.26)
GENDER 1	1.25 (0.13)***	0.19 (0.13)	0.00 (0.12)	-0.30 (0.11)***	-0.82 (0.14)***
AGE	-0.12 (0.07)	0.42 (0.07)***	-0.98 (0.08)***	0.43 (0.06)***	0.50 (0.08)***
SOCEC_LEVEL 2	-0.31 (0.24)	-0.41 (0.25)*	-0.17 (0.26)	0.44 (0.29)	-1.12 (0.21)***
SOCEC_LEVEL 3	-0.28 (0.24)	-0.21 (0.25)	0.29 (0.25)	1.18 (0.28)***	-1.35 (0.22)***
SOCEC_LEVEL 4	-0.48 (0.23)**	-0.47 (0.24)	0.25 (0.24)	1.10 (0.27)***	-1.97 (0.21)***
SOCEC_LEVEL 5	-0.49 (0.28)*	-0.57 (0.28)**	0.02 (0.3)	1.45 (0.29)***	-3.10 (0.35)***
TOT_MEM	0.21 (0.07)***	0.10 (0.07)**	0.11 (0.07)	0.08 (0.06)	-0.17 (0.07)**
CO2_EMISSIONS	-0.40 (0.12)***	0.09 (0.11)	0.20 (0.12)*	0.22 (0.10)**	0.04 (0.12)
RURALITY	-0.48 (0.11)***	0.15 (0.09)*	0.04 (0.09)	0.12 (0.08)	-0.11 (0.10)
MAX_TEMP	-0.02 (0.08)	-0.05 (0.08)	0.16 (0.08)*	0.08 (0.07)	0.10 (0.09)

Note:
Logistic regression coefficients with standard errors in parentheses with Cluster 1 as reference category for the dependent variable; *p<0.2, **p < 0.05, ***p < 0.01

5.3.2 Random Forest

The objective of using Random Forest is to get a better understanding of the relevance of the variables and also to test the impact of the Expansion Factor. It is pertinent to mention that the results from Random Forest were compared to the results of Boosting (from the package "Xgboost" in R). Nevertheless, no significant improvement in terms of accuracy was observed considering the computational demand of the second algorithm mentioned when tuning the parameters and fitting the final model, thus only the results from Random Forest are presented.

To construct the Random Forest model, the same variables from Model A, the one with the best accuracy, are considered, namely REGION, AREA, AGE, GENDER, SOCEC_LEVEL, and TOT_MEM. The observations of the training data are replicated according to the Expansion Factor, resulting in a total of 11,443,585 observations. The test set remains the same as in the previous models, thus 1,166 observations.

Aiming to tackle the class imbalance and also improve the accuracy with this new

expanded sample, 100 Random Forest models are trained with samples of 40,000 random observations. The tuned parameters that yield the lowest classification error are m equal to one, node size equal to 10, and the number of trees equal to 300.

The average out-of-bag accuracy is 51.57%, however, is important to remember that in this case, the training data has replicated observations due to the expansion factor. The average test accuracy resulted in 36.19%, thus slightly lower than the previous models. In Table 5.3.2.1 the test accuracy per cluster can be observed. Compared to the results from the previous models, the accuracy for Cluster 1, 3, and 6 suffered an important drop, whereas the accuracy for Cluster 2 and 5 improved.

Table 5.3.2.1: Confusion Matrix of the Test predictions Random Forest

		Reference					
		1	2	3	4	5	6
Prediction	1	25	4	15	17	12	8
	2	30	58	32	18	33	15
	3	3	1	1	3	0	0
	4	32	12	15	88	29	17
	5	129	34	96	55	207	95
	6	13	6	8	5	7	43
Accuracy		10.78 %	50.44%	0.06%	47.31%	71.88 %	24.16%

In line with the results from the linear models, AGE, GENDER, REGION, and SOCEC_LEVEL appear to be the most important variables (see Table 5.3.2.2: Variable Importance per Cluster derived from Random Forest model), presenting a decrease in the accuracy of 10.86%, 7.36%, 7.13%, and 6.58% respectively when the values of the said variables are permuted. Nonetheless, the importance of these variables varies across Cluster. For instance, in the case of Cluster 1, the most important variables are the region to which the individuals belongs to and the age, whereas the area does not seem relevant. In contrast, for Cluster 2 gender and age are the most relevant variables, followed by the region and total members of the household. On the other hand, the permutation of values of the variables seems similarly relevant when predicting Cluster 3, however, it can be

highlighted that region and age are the most important features, and that area has the largest effect over this cluster. For Cluster 4, age is not only the most important variable, but it is also the one that generates the largest impact on the accuracy when observing the whole panorama, in contrast, the area of living does not appear to influence to a large extent the prediction of this category. When looking at the importance of the variables for Cluster 5, age, gender, and socioeconomic level are the most relevant ones. Lastly, the most important variable for Cluster 6 is the socioeconomic level, which is aligned with the findings in the linear models.

Table 5.3.2.2: Variable Importance per Cluster derived from Random Forest model measured as decrease in the accuracy

Overall	
AGE	0.11
GENDER	0.07
REGION	0.07
SOCEC_LEVEL	0.07
TOT_MEM	0.05
AREA	0.02

Clust. 1	
REGION	0.09
AGE	0.07
TOT_MEM	0.06
SOCEC_LEVEL	0.05
GENDER	0.04
AREA	0.02

Clust. 2	
GENDER	0.17
AGE	0.12
REGION	0.08
TOT_MEM	0.06
SOCEC_LEVEL	0.05
AREA	0.03

Clust. 3	
REGION	0.07
AGE	0.06
TOT_MEM	0.05
GENDER	0.04
SOCEC_LEVEL	0.04
AREA	0.04

Clust. 4	
AGE	0.20
REGION	0.07
TOT_MEM	0.06
SOCEC_LEVEL	0.06
GENDER	0.05
AREA	0.02

Clust. 5	
AGE	0.12
GENDER	0.08
SOCEC_LEVEL	0.08
REGION	0.06
TOT_MEM	0.05
AREA	0.02

Clust. 6	
SOCEC_LEVEL	0.13
REGION	0.06
AGE	0.06
GENDER	0.06
TOT_MEM	0.05
AREA	0.02

5.3.3 Main Findings

Dietary patterns, environmental footprints, and healthiness of diets allowed to classify observations into six clusters that are not only reasonable in terms of size and uniqueness but that present characteristics that go in line with what intuition suggests when observing their average sociodemographic features. Furthermore, both interesting and

useful associations can be found when exploring the clusters and such features through supervised methods. Before highlighting the main findings, it is relevant to mention that the main drawback is that despite using the Expansion Factor to increase the number of observations, there is still a bias towards Cluster 5, the majority class. Notwithstanding, the overall prediction accuracy is almost 38%, thus more than two times chance, which is quite optimistic considering that many more factors can influence the dietary choices of a person. Compelling enough, the prediction accuracy for Cluster 2 is surprisingly good considering it is the smallest of the clusters. This becomes even more interesting when compared to Cluster 3, which turned out to be a category hard to predict. This again sheds a light on the fact that sociodemographic variables are not the only aspect that helps to predict diets.

In general, it was discovered that variables are not necessarily influential in the same magnitude across clusters. As a brief depiction of the clusters, it can be said that Cluster 1 could be considered as average in the sense that it is not strongly inclined to one specific diet nor its sociodemographic characteristics are peculiar when contrasted with the rest of the clusters, furthermore, their environmental footprints are not the lowest nor the highest ones.

On the other hand, Cluster 2 is the one with the largest proportion of people with high environmental footprints and also classified under the category "unhealthy", due to their more frequent consumption of meat and alcohol. Gender, Age, socioeconomic level, and household size are the main drivers to classify this cluster, characterized for comprising more men and young adults, by having a mid-socioeconomic level and by having a positive relationship with the number of total members of the household. Moreover, a negative relationship is observed with CO2_EMISSIONS, suggesting that people from this cluster can be found in less industrialized regions when compared to Cluster 1, but not in extremely rural regions, as indicated by the coefficient for the variable RURALITY.

Cluster 3's footprint indicators are amongst the highest, mostly due to their intake of dairy-derived products and meat. It is one of the clusters that presents a strong association with the geolocation of the individuals. It is regarded by presenting a significant association with people that live in rural areas or even from more rural regions in general, particularly from the regions of the South of Chile. Furthermore, this cluster has a positive relationship with age, thus it comprises older people than the other clusters.

In the case of Cluster 4 and even though their environmental footprint is not the worst, it is still quite high and this group should be of interest because it presents the lowest average age, and this is the most relevant variable for them. The high intakes of processed meats, milk derivatives, and sugar explain the footprint index and also should raise an alarm about this kind of behaviour. Additionally, this cluster presents a positive relationship with higher temperatures and CO2 emissions, suggesting that people live found mostly in North and Centre macro zones in more industrialized regions.

Cluster 5 is the one that presents the lowest environmental impact according to the footprint indices, but it is not clear if it is due to the best reasons. In general people from this cluster consume less meat and more light products, which is something positive. However, it has a low socioeconomic level on average, has the highest average age, the highest Body-mass Index, a larger proportion of women, and, interestingly enough, this cluster presents significant positive associations with the regions of the North of Chile and from rural areas. Moreover, it has a positive relationship with CO2_EMISSIONS, suggesting that these people dwell in regions more industrialized.

Cluster 6 is the one with the second-lowest carbon and water footprint. It comprises people with higher socioeconomic levels, this being the most important variable, and it has the highest proportion of "healthy" individuals. Location does not seem to be a driver for the prediction of this cluster, but it is still worth mentioning that more than half of the cases are from the Metropolitan Region.

6. Conclusions

Overall, interesting and useful outputs for public policies could be derived from this study, summarized in Section 6.1. Among the main findings, distinctive characteristics across clusters can be observed, not only in terms of food intake patterns but also in terms of environmental impact and healthy eating indices. What is more, through supervised methods, significant associations between the prediction of clusters and particular sociodemographic variables were discovered. Lastly, limitations and directions for future research are addressed in Section 6.2.

6.1 Summary and Research Question

Throughout this thesis, different techniques were applied in order to identify dietary patterns, clusters, and driver sociodemographic variables of the population of Chile. The study was based on information gathered from the National Food Consumption Survey 2010, comprising both food-intake related variables and sociodemographic features, a variety of regional variables and water and carbon footprint indices. The dietary patterns were constructed by means of Principal Components Analysis performed over the food intake frequency of individuals, discovering five principal components or patterns named "Diverse and light foods," "Low intake and light foods," "Southern traditional and animal products intensive," "Meat and alcohol inclined," and "Sugar intensive," according to their foods' loadings. Subsequently, the dietary patterns in addition to a healthy eating index, carbon footprint index, and water footprint index, were the foundation to find representative clusters with the k-Prototypes algorithm. This way, six clusters were discovered, presenting distinctions when it comes to food choices, environmental impact, and healthiness. Moreover, differences across sociodemographic and regional variables could also be observed.

More specifically, Cluster 1 could be considered "average" in the sense that it presents a rather diverse diet, mid-level footprint, and health indices, and regarding sociodemographic variables, it could be emphasised the fact that this is the second cluster with the highest proportion of people that belong to rural areas. Cluster 2 presents the

worst environmental footprint and health indices, due to a diet mostly concentrated around alcohol and meats. This later cluster is characterized by comprising more males in comparison to the other clusters, by being mostly young adults and having a larger number of members in the household. On the other hand, Cluster 3 is the one with the second-highest environmental footprint, mostly explained for high intakes of different types of meats, dairy products, and eggs, thus strongly animal-based. This cluster is rather interesting because it is mostly concentrated in the southern regions of Chile and age has a positive relationship with it. Cluster 4 has moderately high water and carbon footprints. The dietary choices are mostly inclined towards sweets, thus sugar intensive, processed meats and milk derivatives, and their main sociodemographic characteristic are that teenagers and young adults take part in it, principally from more industrialized regions. When it comes to Cluster 5, it could be said that it has the least negative environmental impact, comprising people from lower socioeconomic levels and principally from rural areas of the North of Chile, however it still includes people from other regions. Lastly, Cluster 6 does not present a high environmental impact, its diet is rather diverse and inclined to light foods and it is composed mostly by people from higher socioeconomic levels and the Metropolitan region, these two being significant predictor variables.

Going back to research question, "which are relevant groups and the sociodemographic variables that affect them on which the Chilean government should focus in order to comply with the Sustainable Development Goal 12?", it could be argued that the most relevant groups and on which focus should be put are Cluster 2, Cluster 3 and Cluster 4. The three of them are different and could be tackled by targeting people with particular characteristics. For instance, a good way to generate awareness for Cluster 2 could be in universities and workplaces at a national level, to create more conscious eating habits when it comes to meat and alcohol consumption since this group seems to have a rather unhealthy diet and a fast lifestyle. In the case of Cluster 3, attention should be centred on regions of the South, principally for older people from the countryside, raising cognizance of the environmental and health impacts of choosing meat and dairy as main

components of their daily meals. Advantage could be taken of the fact that this cluster is closer to products produced locally, thus more environmentally-friendly, and that their intake of fruits, vegetables, and grains is already high enough to achieve a balanced diet. Finally, Cluster 4, characterized by being the youngest and by presenting poor eating habits, could be targeted in schools through modifying the national program of education in nutrition, to put extra focus on the negative impact of processed meats over people's health and for the planet, in addition to more consciousness of the effect of sugar on the body.

These suggestions should be taken merely as a starting point to focus the Chilean government's efforts on accelerating the implementation of public policies in aims to comply with the Sustainable Development Goal 12, namely sustainable consumption. Notwithstanding, it is understood that further actions should be taken at a national scale to generate a broader impact.

6.2 Limitations and Further Improvements

In spite of the discovery of interesting and relevant relationships between the independent variables and the clusters, the prediction capability of the model could still be improved if additional features were to be incorporated, hopefully, available in the database from the national Census or other sources so that the model could be applied over the population, namely field of work, ethnic origins, religion, the closeness of house to markets, marital status, health conditions, level of study, and so on. Furthermore, extra variables describing regional features could also be included, such as better detail of the economic activity, types of natural resources, among others. Nonetheless, it is important to acknowledge the fact that, ultimately, diets are also influenced by preferences, trends, traditions, and several other external influences, so the predictive accuracy can only be improved to some extent.

Additionally, sample size plays a major role in models that comprise various categorical variables, thus it would be expected for the model to return more accurate results if the sample was larger. Eventually, the inclusion of comparable data from other countries could help to expand the sample size and the variety of observations. However, it is important to keep in mind that due to the geographic characteristics and economic activity of Chile, it is not recommended to extrapolate the results of the current model to other countries.

Another important element to take into consideration is the fact that the population in Chile has changed since 2010. The number of immigrants has increased, meaning that new customs are being brought to the country. Also, vegetarianism/veganism is still trending and such diets or life-style have more adepts than 10 years ago. Thus, even though patterns do not change so fast, using data more up to date could eventually disclose new dietary patterns, new clusters, and, of course, more recent results.

The model could be further complexified in terms of precision of the dietary patterns and its environmental impact if other food intake-related variables more detailed are included, such as consumption over several months, the time of the day when the food is usually consumed, the origin of the food (homemade, pre-made-, restaurant, and so on.), brands, among others.

Lastly, it could be said that there is still space for improvements even though this study already provides a starting point for the development of public policies focused on raising awareness about sustainable and healthy food consumption. However, ultimately the need for larger samples and more detailed variables will depend on the objective of the analysis.

References

Ambrosini, GL., Fritschi, L., Hubert de Klerk, N., Mackerras, D. & Leavy, J. (May, 2018). Dietary Patterns Identified Using Factor Analysis and Prostate Cancer Risk: A Case Control Study in Western Australia. *Annals of Epidemiology*; 18:364-370. doi: https://doi.org/10.1016/j.annepidem.2007.11.010

BCFN Foundation (2015). Recommendations for a Sustainable Diet. Double Pyramid (2015): 67-69

Poore, J. & Nemecek, T. (June, 2018). Reducing food's environmental impacts through producers and consumers. *Science*; 360(6392): 987-992. doi: https://doi.org/10.1126/science.aaq0216

Biosbroek, S., Verschuren, VMM., van der Schouw, YT., Sluijs, I., Boer, JMA., & Temme, EHM. (2018). Identification of data-driven Dutch dietary patterns that benefit the environment and are healthy. *Climatic Change*; 147:571-583. doi: https://doi.org/10.1007/s10584-018-2153-y

BMJ. (2018). Role of government policy in nutrition - barriers to and opportunities for healthier eating; 361:k2426. Retrieved June 22nd, 2020, from https://www.bmj.com/content/361/bmj.k2426

Curran, M.A. (June, 2008). Development of life cycle assessment methodology: a focus on co-product allocation. Erasmus University Rotterdam. Retrieved May 3rd, 2020, from http://hdl.handle.net/1765/12679

Devlin, UM., McNulty, BA., Nugent, AP. & Gibney, MJ. (August, 2012). The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy misreporting. *Proceeding of the Nutrition Society*; 71:599-609. doi: https://doi.org/10.1017/s0029665112000729

Fernandez, P. (March, 2017). ¿De dónde viene la carne que consumimos en Chile? [Where does the meat we consume in Chile come from?]. *24 horas*. Retrieved May 18th, 2020, from https://www.24horas.cl/nacional/de-donde-viene-la-carne-que-consumimos-en-chile--2332654

Food and Agricultural Organization of the United Nations & Pan American Health Organization (2017). Aprobación de la nueva Ley de Alimentos en Chile: Resumen del proceso [Approval of the new Food Law in Chile: Summary of the process]. Retrieved May 3rd, 2020, from http://www.fao.org/3/a-i7692s.pdf

González, A. Frostell, B. & Carlsson-Kanyama, A. (July, 2011). Protein efficiency per unit energy and per unit greenhouse gas emissions: Potential contribution of diet

choices to climate change mitigation. *Food Policy*; 36: 562-570. doi: https://doi.org/10.1016/j.foodpol.2011.07.003

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *BioMetrics*; 27: p. 857-874. doi: https://doi.org/10.2307/2528823

Huseinovic, E., Hörnell, A., Johansson, I., Esberg, A., Lindahl, B. & Winkvist, A. (2019). Changes in food intake patterns during 2000-2007 and 2008-2016 in the population-based Northern Sweden Diet Database. *Nutrition Journal, BMC*; 18, article 36. doi: https://doi.org/10.1186/s12937-019-0464-0

Hammerling, U., Freyhult, E., Edberg, A., Sand, S., Fagt, S., Knudsen, VK., Andersen, LF., Lindroos, AK., Soeria-Atmadja, D. & Gustafsson, MG. (July, 2014). Identifying Food Consumption Patterns among Young Consumers by Unsupervised and Supervised Multivariate Data Analysis. *European Journal of Nutrition and Food Safety*; 4(4):392-403. doi: https://doi.org/10.9734/ejnfs/2014/9082

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore; p. 21-34.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*; 2: 283-304.

Institute of Agricultural Research of Chile (2013). Determinación de la huella de agua y estrategias de manejo de recursos hídricos [Determination of the Water Footprint and water resource management strategies]; (50). Retrieved May 18th, 2020, from http://biblioteca.inia.cl/medios/biblioteca/serieactas/NR38988.pdf

Institute of Nutrition and Food Technology of University of Chile (June, 2019). A tres años de Ley Etiquetado: Cambio en la composición de productos y en los hábitos de compra [Three years after the Labeling Law: Change in the composition of products and purchasing habits]. Retrieved May 3rd, 2020, from https://inta.cl/a-tres-anos-de-ley-etiquetado-cambio-en-la-composicion-de-productos-y-en-los-habitos-de-compra/

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2017). An Introduction to Statistical Learning with Applications in R. Chapter 10: 374-379. doi: https://doi.org/10.1111/insr.12051_19

Kastorini, C., Papadakis, G., Milionis, HJ., Kalantzi, K., Puddud, P., Nikolaoue, V., Vemmos, KN. , Goudevenos, JA. & Panagiotakos, DB. (August, 2013). Comparative analysis of a-priori and a-posteriori dietary patterns using state-of-the-

art classification algorithms: A case/case-control study. *Artificial Intelligence in Medicine*; 59:175-183. doi: https://doi.org/10.1016/j.artmed.2013.08.005

Kramer, GFH., Tyszler, M., van't Veer, P. & Blonk, H. (March, 2017). Decreasing the overall environmental impact of the Dutch diet: how to find healthy and sustainable diets with limited changes. *Public Health Nutrition*; 20(9): 1699-1709. doi: https://doi.org/10.1017/s1368980017000349

Krebs-Smith, SM., Pannucci, TRE., Subar, AF., Kirkpatrick, SI., Lerman, JL., Tooze, JA., Wilson, MM. & Reedy, J. (September, 2018). Update of the Healthy Eating Index: HEI-2015. *Journal of the Academy of Nutrition and Dietetics*; 118(9): 1591-1602. doi: https://doi.org/10.1016/j.jand.2018.05.021

Lazaroua, C., Karaolis, M., Matalas, A. & Panagiotakos, DB. (December, 2011). Dietary patterns analysis using data mining method. An application to data from the CYKIDS study. *Computer Methods and Programs in Biomedicine*; 108:706-714. doi: https://doi.org/10.1016/j.cmpb.2011.12.011

Manuka (2019). Reporte de Sustentabilidad 2019 [Sustainability Report 2019]. Retrieved May 18th, 2020, from http://www.manuka.cl/web/wp-content/uploads/2019/10/Reporte-Espa%C3%B1ol-Final-2019-09-24.pdf.

Marchetti, P. (June, 2018). Encuesta de Presupuestos Familiares: Gasto promedio mensual de hogares chilenos sobrepasa el millón de pesos. *Emol*. Retrieved March 15th, 2020, from https://www.emol.com/noticias/Economia/2018/06/25/911091/Encuesta-de-Presupuestos-Familiares-Gasto-promedio-mensual-de-hogares-chilenos-sobrepasa-el-millon-de-pesos.html

McCann, SE., Marshall, JR., Brasure, JR., Graham, S. & Freudenheim JL. (March, 2001). Analysis of patterns of food intake in nutritional epidemiology: food classification in principal components analysis and the subsequent impact on estimates for endometrial cancer. *Public Health Nutrition*; 4(5): 989-997. doi: https://doi.org/10.1079/phn2001168

Ministry of Environment of Chile (2017). Plan de Acción Nacional de Consumo y Producción Sustentables 2017-2022 [National Plan of Action for Sustainable Consumption and Production 2017-2022]. Santiago, Chile. Retrieved May 3rd, 2020, from https://mma.gob.cl/wp-content/uploads/2017/11/PLAN-NACIONAL-DE-ACCION-CPS-2017-2020.pdf

Ministry of Health of Chile (May, 2013). Informe final: Estudio para la revisión y actualización de las guías alimentarias para la población chilena [Final report: Study for the revision and update of the dietary guidelines for the Chilean population].

Resolución Exenta N°260; p.134. Santiago, Chile. Retrieved May 3rd, 2020, from https://www.minsal.cl/portal/url/item/dde0bc471a56a001e040010165012224.pdf

Neumark-Sztainer, D., Hannan, PJ., Story, M., Croll, J., & Perry, C. (March, 2003). Family meal patterns: Associations with sociodemographic characteristics and improved dietary intake among adolescents. *Journal of the American Dietetic Association*; 103(3): 317-322. doi: https://doi.org/10.1053/jada.2003.50048

Nishi, N., Horikawa, C. & Murayama, N. (January, 2017). Characteristics of food group intake by household income in the National Health and Nutrition Survey, Japan. *Asia Pacific Journal of Clinical Nutrition*; 26(1):156-159. doi: https://doi.org/10.6133/apjcn.102015.15

Norte Navarro, AI. & Ortiz Moncada, R. (March-April, 2011). Spanish diet quality according to the Healthy Eating Index. *Nutrición Hospitalaria*; 26(2): 1699-5198 online version & 0212-1611 printed version. Retrieved May 3rd, 2020, from http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-16112011000200014

OECD (April, 2013). Glossary of Statistical Terms, Carbon Dioxide Equivalent. Statistics Portal. Retrieved May 10th, 2020, from https://stats.oecd.org/glossary/detail.asp?ID=285

Office of Agricultural Studies and Policies of Chile (August, 2018). Boletín de la leche: producción, recepción, precios y comercio exterior [Milk bulletin: production, reception, prices and foreign trade]. Retrieved May 18th, 2020, from https://www.odepa.gob.cl/wp-content/uploads/2018/08/Informe-lacteo-ago-2018.pdf

Ortega, RM. & Requejo, AM. (2000). Introducción a la Nutrición Clínica. Nutriguía, Manual de Nutrición Clínica en Atención Primaria; Chapter 9: 85-93

Padmadas, SS., Dias, JG. & Willekens, FJ. (January, 2007). Disentangling women's responses on complex dietary intake patterns from an Indian cross-sectional survey: A latent class analysis. Cambridge University Press. *Public Health Nutrition*; 9(2):204-11. doi: https://doi.org/10.1079/phn2005842

Panagiotakos, DB., Pitsavos, C. & Stefanadis, C. (December, 2006). Dietary patterns: A Mediterranean diet score and its relation to clinical and biological markers of cardiovascular disease risk. *Nutrition, Metabolism and Cardiovascular Diseases*; 16(8): 559-568. doi: https://doi.org/10.1016/j.numecd.2005.08.006

Sala, S. & Castellani, V. (December, 2019). The consumer footprint: Monitoring sustainable development goal 12 with process-based life cycle assessment. *Journal of*

Cleaner Production; 240(2019): 118050. doi: https://doi.org/10.1016/j.jclepro.2019.118050

Stehfest, E., Bouwman, L., van Vuuren, DP., den Elzen, MGJ., Eickhout, B. & Kabat, P. (February, 2009). Climate Benefits of Changing Diets. *Climatic Change*; 95: 83-102. doi: https://doi.org/10.1007/s10584-008-9534-6

Szepannek, G. (2018). clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *The R Journal*; 10(2): p. 200-208. doi: https://doi.org/10.32614/rj-2018-048

Thinsungnoena, T., Kaoungkub, N.,Pongsakorn, D., Kerdprasopb, K. & Kerdprasopb, N. (2015). The Clustering Validity with Silhouette and Sum of Squared Errors. *Proceedings of the 3rd International Conference on Industrial Application Engineering.* doi: https://doi.org/10.12792/iciae2015.012

Thorpe, MG., Milte, CM., Crawford, D. & McNaughton, SA. (February, 2016). A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians. *International Journal of Behavioral Nutrition and Physical Activity*; 13(30). doi: https://doi.org/10.1186/s12966-016-0353-2

Tseng, M., Breslow, RA., DeVellis, RF. & Ziegler, RG. (January, 2004). Dietary Patterns and Prostate Cancer Risk in the National Health and Nutrition Examination Survey Epidemiological Follow-up Study Cohort. *Cancer Epidemiology, Biomarkers & Prevention*; 13: 71-77. doi: https://doi.org/10.1158/1055-9965.epi-03-0076

Tucker, K. (April, 2010). Dietary patterns, approaches and multicultural perspective. *Applied psychology nutrition and metabolism*; 35(2):211-8. doi: https://doi.org/10.1139/h10-010

UN Global Compact (November, 2015). Ecobase Alimentos: la primera calculadora ambiental del país para productos de exportación [Food Ecobase: the country's first environmental calculator for export products]. Retrieved May 3rd, 2020, from https://pactoglobal.cl/2015/ecobase-alimentos-la-primera-calculadora-ambiental-del-pais-para-productos-de-exportacion/

United Nations (2020). Goal 12: Sustainable consumption and production. Retrieved February 25th, 2020, from https://www.un.org/sustainabledevelopment/sustainable-consumption-production/

United Nations (2020). The Sustainable Development Agenda. Retrieved February 25th, 2020, from https://www.un.org/sustainabledevelopment/development-agenda/

University of Chile (2014). Informe final: Encuesta Nacional de Consumo Alimentario. Facultad de Medicina y Facultad de Economía y Negocios, Universidad de Chile para

el MINSAL (Ministerio de Salud) [Final report: National Survey of Food Consumption. Faculty of Medicine and Faculty of Business and Economics, University of Chile for the Health Ministry]. Santiago, Chile. Retrieved April 7th, 2020, from https://www.minsal.cl/sites/default/files/ENCA-INFORME_FINAL.pdf

U.S. Department of Health and Human Services and U.S. Department of Agriculture (December, 2015). 2015-2020 Dietary Guidelines for Americans. 8th Edition. Retrieved April 10th, 2020, from http://health.gov/dietaryguidelines/2015/guidelines/

Appendix A: Healthy Eating Index for the Spanish population

Table A.1: Scoring system of the Healthy Eating Index for the Spanish population

Variables	Criteria for max. score of 10	Criteria for max. score of 7.5	Criteria for max. score of 5	Criteria for max. score of 2.5	Criteria for min. score of 0
Daily consumption					
1. Cereals and derivatives	Daily consumption	3 ≥ times per week	1 or 2 times per week	Less than 1 time per week	Never or almost never
2. Greens and vegetables	Daily consumption	3 ≥ times per week	1 or 2 times per week	Less than 1 time per week	Never or almost never
3. Fruits	Daily consumption	3 ≥ times per week	1 or 2 times per week	Less than 1 time per week	Never or almost never
4. Milk and derivatives	Daily consumption	3 ≥ times per week	1 or 2 times per week	Less than 1 time per week	Never or almost never
Weekly consumption					
5. Meats	1 - 2 times per week	3 ≥ times per week	< 1 time per week	Daily consumption	Never or almost never
6. Legumes	1 - 2 times per week	3 ≥ times per week	< 1 time per week	Daily consumption	Never or almost never
Occasional consumption					
7. Sausages and cold cuts	Never or almost never	< 1 time per week	1 - 2 times per week	3 ≥ times per week	Daily consumption
8. Sweets	Never or almost never	< 1 time per week	1 - 2 times per week	3 ≥ times per week	Daily consumption
9. Soft drinks with sugar	Never or almost never	< 1 time per week	1 - 2 times per week	3 ≥ times per week	Daily consumption
10. Variety	2 points if the person complies to every daily recommendations and 1 point if complies to every weekly recommendations				

Note:

Table adapted from "Spanish diet quality according to the Healthy Eating Index" by Norte Navarro, A.I. and Ortiz Moncada, R., 2011. Nutrición Hospitalaria; 26(2): 1699-5198 on-line version & 0212-1611 printed version.

Appendix B: Variables extracted from the National Food Consumption Survey

Table B.1: Variables comprised in the National Food Consumption Survey data set and its corresponding descriptions

Variable	Description
FOLIO	Categorical variable indicating the unique identifier of the subject interviewed. There are 5,120 folios.
PRODUCT_CODE	Categorical variable indicating the unique code assigned to the food item. There are 457 food items, thus 457 codes.
PRODUCT_NAME	Categorical variable indicating the name of the food item. There are 457 food items.
CONSUME	Binary variable indicating whether or not the product is usually consumed in a month, 0 for "no" and 1 for "yes".
FREQ_MONTH	Integer numerical variable indicating the frequency of consumption of a food item.
CONS_MONTH	Continuous numerical variable indicating the consumption in grams (in the case of solids) or millilitres (in the case of liquids) of food items.
GENDER	Binary variable indicating the gender of the subject, 0 for women and 1 for men.
AGE	Numerical variable indicating the age of the subject.
AGE_GROUP	Categorical variable indicating the age group to which the subject belongs to. There are 7 levels: <6 , 6-13, 14-18, 19-29, 30-49, 50-64 and >=65 years old.
BMI	Numerical variable indicating the Body Mass Index of the subject.
HEIS	Categorical variable indicating the Health Eating Index (from Spain) of the subject. There are three categories: 1 for "healthy", 2 for "requires changes" and 3 "unhealthy".
GLOBIND	Categorical variable indicating the number of dietary recommendations (from the Health Ministry of Chile) complied by the subject. There are three categories: 1 for "satisfactory compliance", 2 for "partial compliance" and 3 for "no compliance".
SOCEC_LEVEL	Categorical variable indicating the socio-economical level of the subject. There are five levels, from 1 to 5 that represent High, Medium High, Medium, Medium Low and Low respectively.
MACRO_ZONE	Categorical variable indicating the macro zone to which the household belongs to. There are five macro zones: North, North Centre, South Centre, South, and Metropolitan.
REGION	Categorical variable indicating the Region to which the household belongs to. There are 15 Regions.
COMMUNE	Categorical variable indicating the commune to which the household be-longs to, in Spanish corresponds to "Comuna". There are 111 Cities.
EXP_F	Numerical variable indicating the Expansion Factor corresponding to the subject.
AREA	Numerical variable indicating whether the individual belongs to a rural area (0) or urban area (1).
TOT_MEM	Numerical variable indicating the total number of people that live in the household with the subject.

Appendix C: National Food Consumption Survey general statistics

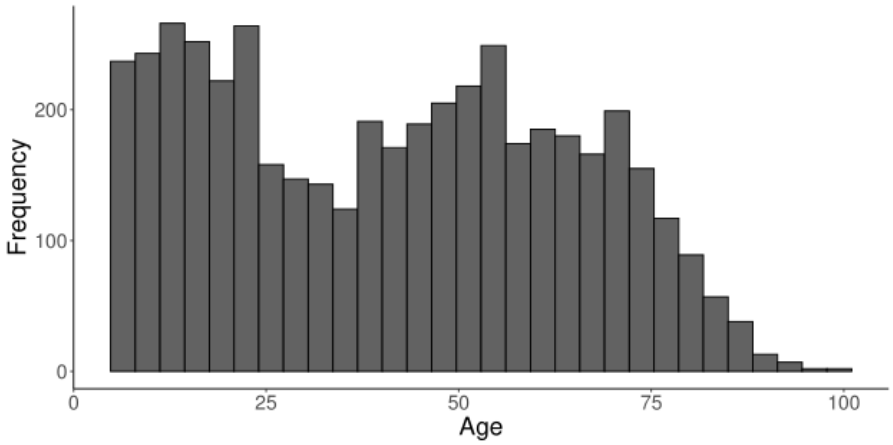

Figure C.1: Age distribution

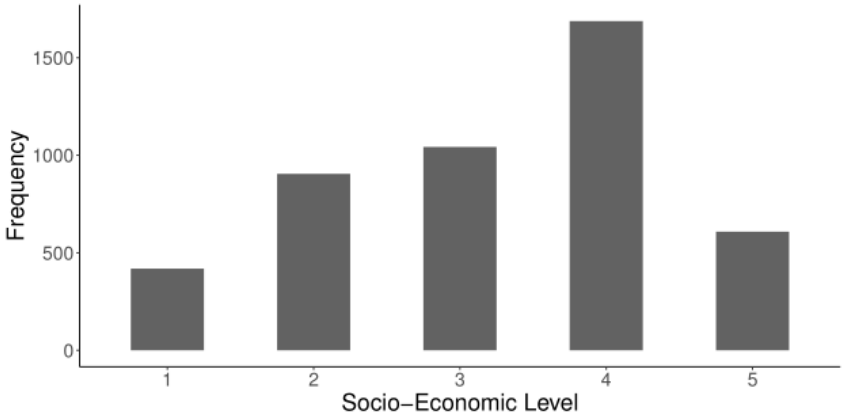

Figure C.2: Socioeconomic level distribution

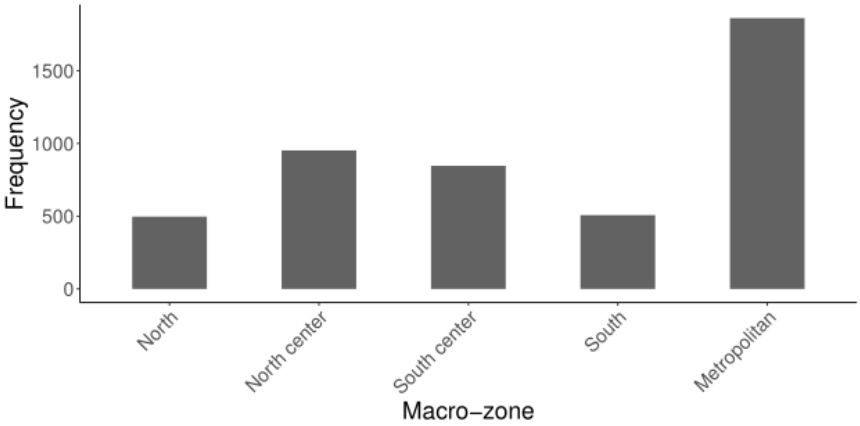

Figure C.3: Macro zone distribution

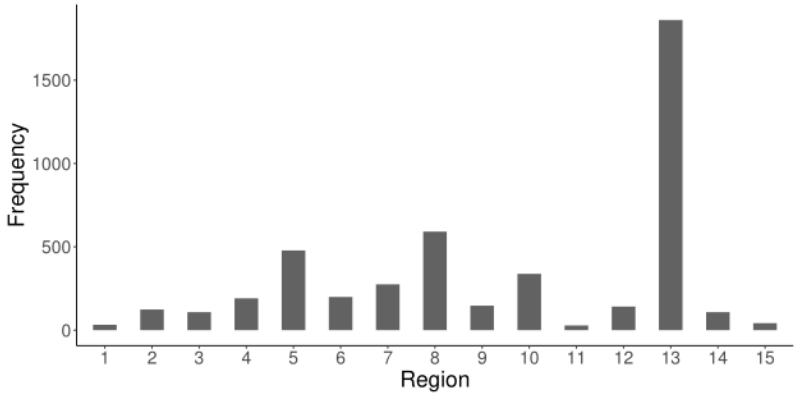

Figure C.4: Region Frequency Distribution

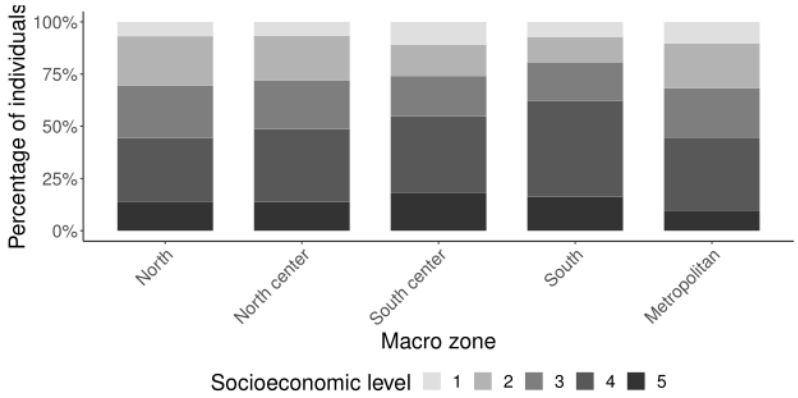

Figure C.5: Socioeconomic level Frequency Distribution per Macro Zone

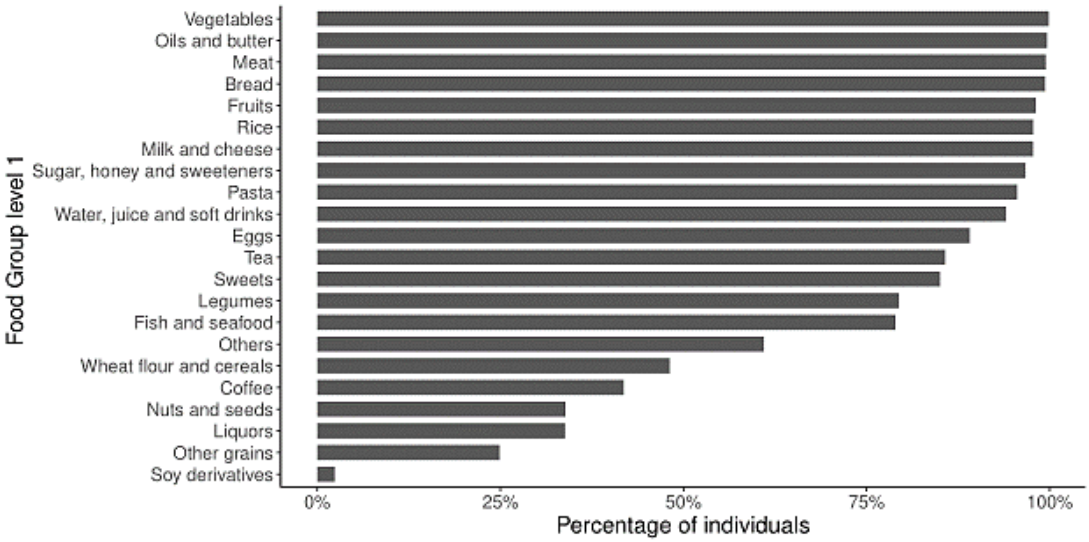

Figure C.6: Percentage of consumers per food group (aggregation level 1)

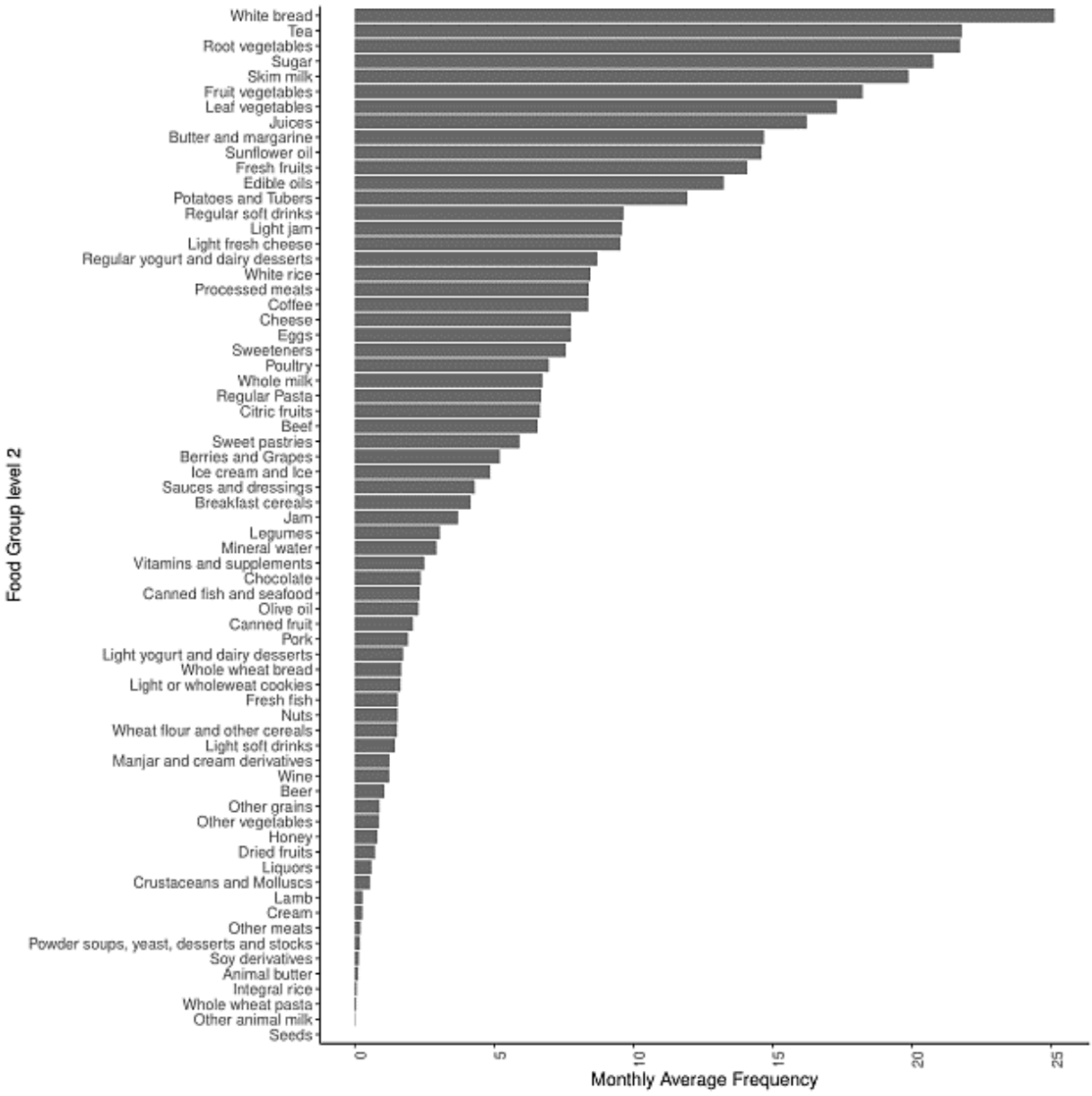

Figure C.7: Intake frequency of grouped food (aggregation level 2)

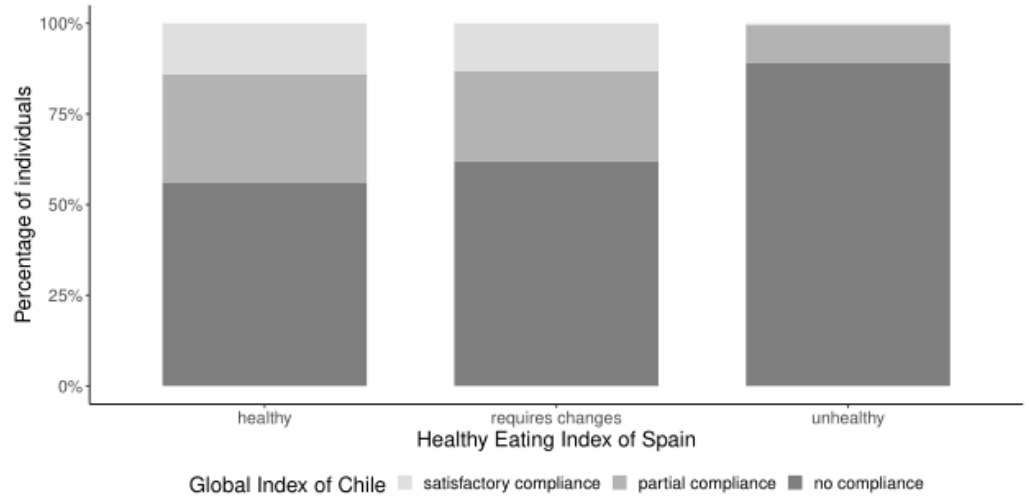

Figure C.8: Healthy Eating index of Spain crossed with Global Index of Chile

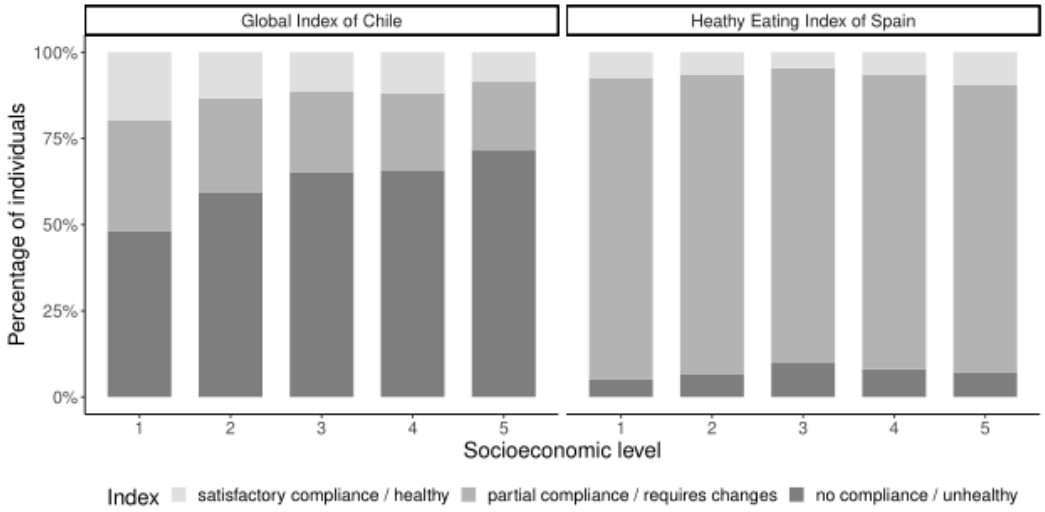

Figure C.9: Healthy Eating index of Chile and Spain across socioeconomic levels

Table C.1: Food groups aggregation level 1 and level 2

Food Groups Level 2	Food groups Level 1	Food Groups Level 2	Food groups Level 1
Eggs	Eggs	Fruit vegetables	Bread
Beef	Meat	Leaf vegetables	Vegetables
Lamb	Meat	Other vegetables	Vegetables
Other meats	Meat	Potatoes and Tubers	Vegetables
Pork	Meat	Root vegetables	Vegetables
Poultry	Meat	Fresh fruits	Fruits
Processed meats	Meat	Berries and Grapes	Fruits
Canned fish and seafood	Fish and seafood	Canned fruit	Fruits
Crustaceans and Molluscs	Fish and seafood	Citric fruits	Fruits
Fresh fish	Fish and seafood	Dried fruits	Fruits
Cheese	Milk and cheese	White bread	Bread
Cream	Milk and cheese	Whole wheat bread	Bread
Light fresh cheese	Milk and cheese	Regular Pasta	Pasta
Light yogurt and dairy desserts	Milk and cheese	Whole wheat pasta	Pasta
Manjar and cream derivatives	Milk and cheese	Integral rice	Rice
Other animal milk	Milk and cheese	White rice	Rice
Regular yogurt and dairy desserts	Milk and cheese	Honey	Sugar, honey and sweeteners
Skim milk	Milk and cheese	Jam	Sugar, honey and sweeteners
Whole milk	Milk and cheese	Light jam	Sugar, honey and sweeteners
Nuts	Nuts and seeds	Sugar	Sugar, honey and sweeteners
Seeds	Nuts and seeds	Sweeteners	Sugar, honey and sweeteners
Animal butter	Oils and butter	Chocolate	Sweets
Butter and margarine	Oils and butter	Ice cream and Ice	Sweets
Edible oils	Oils and butter	Light or whole wheat cookies	Sweets
Olive oil	Oils and butter	Sweet pastries	Sweets
Sunflower oil	Oils and butter	Soy derivatives	Soy derivatives
Other grains	Other grains	Coffee	Coffee
Legumes	Legumes	Tea	Tea

Appendix D: Regional correspondence and features

Table D.1: Region's names, numbers and characteristics

Name of the Region	Region Number	CO2_emissions	Rurality (%)	Agri_surface (ha)	Max_Temp (Celsius)	Agri_people (%)
Tarapacá	I	2,018.80	6.20	53,177.70	24.00	4.12
Antofagasta	II	22,299.70	5.90	2,412.40	24.00	2.09
Atacama	III	6,830.00	8.96	19,734.70	26.00	6.26
Coquimbo	IV	2,079.10	18.80	152,136.70	19.00	11.31
Valparaíso	V	16,937.40	8.99	154,988.80	19.00	8.11
Libertador General Bernardo O'Higgins	VI	2,322.60	53.00	363,835.20	26.00	22.28
Maule	VII	1,472.30	26.80	761,981.20	26.00	23.48
Biobío	VIII	13,476.10	21.00	1,209,158.00	21.00	0.12
La Araucanía	IX	-4,680.20	29.10	916,992.00	22.00	11.59
Los Lagos	X	-7,036.00	26.40	202,085.50	18.00	14.98
Aysén del General Carlos Ibáñez del Campo	XI	- 19,741.00	20.40	55,500.90	15.00	12.56
Magallanes y de la Antártica Chilena	XII	-7,359.90	8.10	6,768.30	12.00	7.52
Metropolitana de Santiago	XIII	22,689.70	3.70	149,991.30	27.00	2.10
Los Ríos	XIV	-5,826.30	28.30	350,824.20	20.00	16.11
Arica y Parinacota	XV	700.20	8.33	6,693.40	24.00	8.53

Note:
Carbon footprint is measured in kilotons of carbon dioxide equivalent ($kTCO_2eq$).

Appendix E: Food environmental footprint

Table E.1: Estimated carbon and water footprint

Food category	Carbon footprint	Water footprint
Beef	33,265	23,449
Poultry	4,123	4,805
Pork	5,619	4,009
Lamb	24,400	757
Cheese	15,225	6,260
Fish	3,720	36
Crustaceans	6,872	-
Eggs	2,908	2,282
Fresh Milk	1,545	762
Powder Milk	18,758	502
Butter	8,305	5,555
Yogurt	1,214	1,050
Bread	1,262	1,090
Pasta	1,900	1,710
Rice	3,815	1,792
Breakfast cereals	3,420	920
Cereals, grains and seeds	1,300	920
Olive oil	4,508	9,650
Sunflower oil	3,600	3,919
Sweet pastries	1,610	2,075
Sweets	2,235	2,410
Chocolate	18,700	2,400
Sweeteners	1,000	-
Honey	1,000	-
Sugar and jam	2,100	200
Margarine	1,360	1,325
Fruit vegetables	815	335
Tomatoes	1,400	150
Avocado	1,576	1,100
Corn	815	500
Root and leaf vegetables	400	335
Potatoes	805	555
Fruit	338	930
Apples	316	112
Banana	700	86
Plum	319	712
Peach	138	250
Citrus fruit	338	600
Berries and Grapes	338	532
Canned fruit	2,685	103
Dried fruit	2,827	3,532
Nuts	300	3,532
Legumes	1,648	2,710
Wine and Liquors	1,594	17
Beer	1,100	120
Soft drinks	900	570
Soy derivatives	1,900	165
Coffee	16,500	120

Note:

Carbon footprint is measured in grams of carbon dioxide equivalent (gCO_2eq) per kilogram or litre of food, and water footprint is measured in litres of water resources used per kilogram or litre of food.

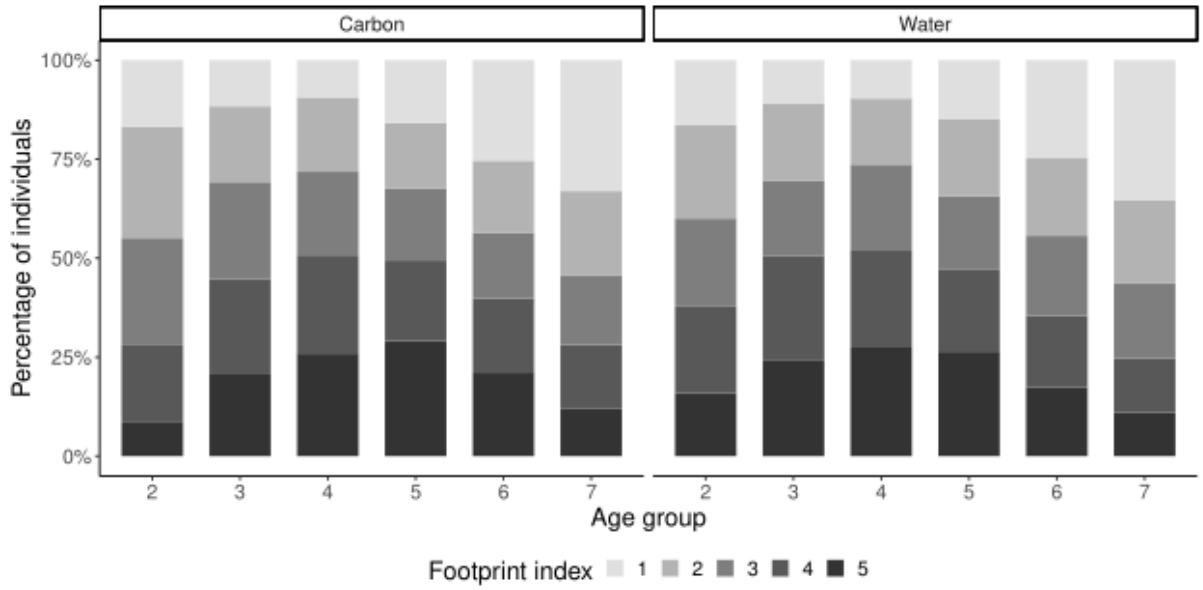

Figure E.1: Carbon and Water footprint across age groups

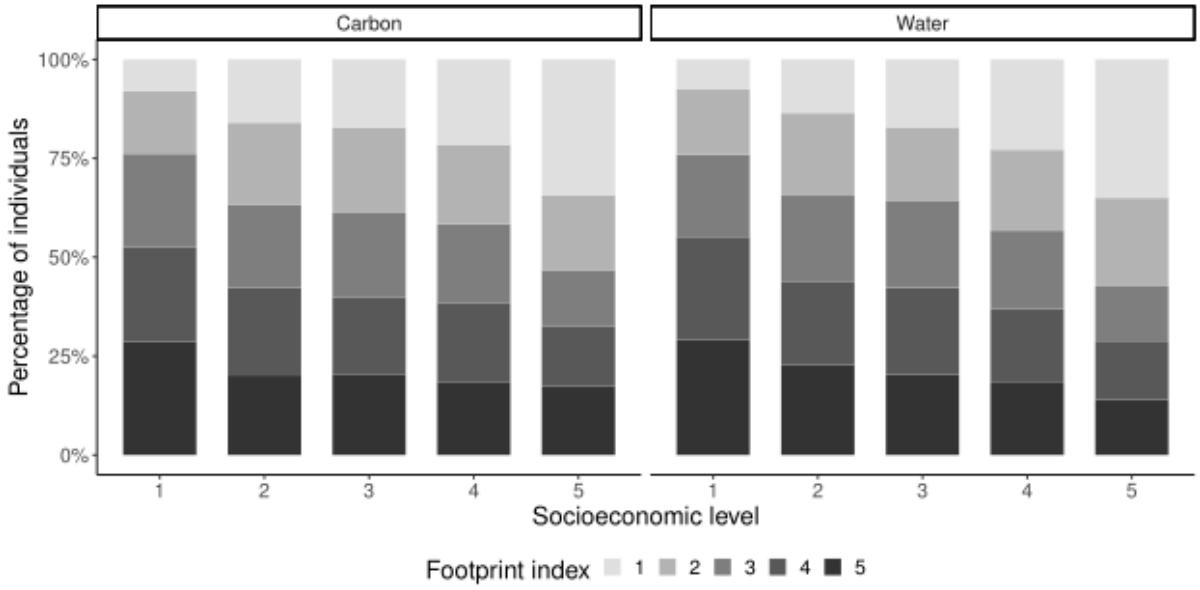

Figure E.2: Carbon and Water footprint across socioeconomic levels

Appendix F: Principal Components loadings

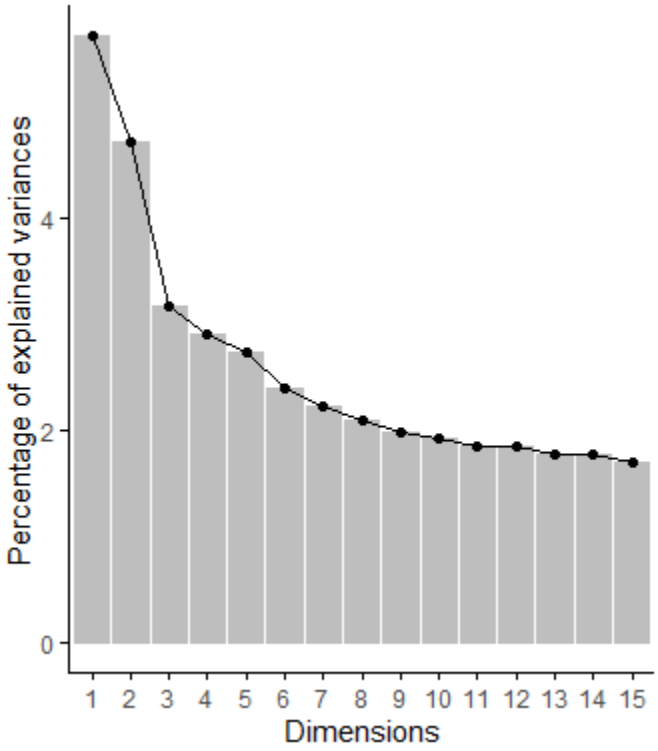

Figure F.1: Scree plot of the cumulative percentage of explained variance for Principal Components Analysis

Table F.1: Principal component loadings

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Sunflower oil	0.06	0.02	0.09	0.61	0.14
Olive oil	0.22	-0.01	-0.06	0.06	-0.01
Edible oils	-0.12	-0.05	-0.06	-0.62	-0.15
Mineral water	0.13	0.01	-0.07	-0.03	-0.13
Liquors	0.02	-0.07	-0.18	0.14	-0.31
White rice	-0.08	-0.09	-0.01	0.03	-0.20
Integral rice	0.06	0.00	0.00	0.00	0.00
Sugar	-0.30	-0.15	0.13	0.08	0.03
Regular soft drinks	-0.17	-0.19	-0.12	0.01	-0.09
Light soft drinks	0.16	0.06	-0.13	0.00	-0.06
Berries and Grapes	0.14	-0.12	0.10	-0.08	0.08
Coffee	0.05	-0.07	-0.03	0.17	-0.11
Poultry	0.07	-0.10	0.00	-0.03	-0.12
Pork	-0.07	-0.14	0.03	0.02	-0.12
Lamb	-0.01	-0.07	0.14	0.10	-0.06
Beef	0.02	-0.17	-0.04	0.04	-0.13
Processed meats	0.00	-0.25	-0.19	0.01	-0.10
Breakfast cereals	0.14	-0.12	-0.07	-0.02	0.27
Beer	-0.01	-0.09	-0.12	0.16	-0.34
Chocolate	0.00	-0.21	-0.21	0.01	0.11
Citric fruits	0.15	-0.11	0.08	0.01	0.07

Table F.1: Principal component loadings
(continued)

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Cream	0.04	-0.11	-0.07	0.03	-0.09
Crustaceans and Molluscs	0.06	-0.10	0.04	0.03	-0.21
Soy derivatives	0.06	-0.04	-0.04	0.02	0.03
Sweeteners	0.30	0.16	-0.04	-0.08	-0.07
Canned fruit	0.16	-0.13	0.04	0.00	-0.05
Fresh fruits	0.21	-0.15	0.11	-0.05	0.08
Dried fruits	0.09	-0.10	0.06	0.00	0.03
Nuts	0.17	-0.11	-0.01	-0.02	-0.01
Light or whole wheat cookies	0.21	0.06	-0.03	-0.05	0.00
Wheat flour and other cereals	-0.01	-0.05	0.28	0.03	-0.02
Ice cream and Ice	-0.03	-0.26	-0.18	-0.09	0.17
Fruit vegetables	0.17	-0.13	0.21	-0.16	-0.11
Root vegetables	0.12	-0.11	0.30	-0.09	-0.12
Leaf vegetables	0.19	-0.12	0.25	-0.08	-0.06
Eggs	0.00	-0.19	0.03	0.01	-0.03
Juices	0.05	-0.16	-0.08	-0.06	0.06
Whole milk	-0.06	-0.17	0.04	0.00	0.23
Skim milk	0.20	0.05	-0.05	0.01	0.06
Other animal milk	-0.01	0.00	0.04	0.03	-0.03
Legumes	0.02	-0.08	0.23	0.08	0.06
Manjar and cream derivatives	0.00	-0.17	0.02	-0.01	0.13
Animal butter	-0.01	-0.03	0.07	0.04	-0.06
Butter and margarine	-0.02	-0.16	-0.01	0.00	-0.01
Sweet pastries	-0.03	-0.24	-0.15	-0.07	0.18
Jam	0.02	-0.13	0.14	0.03	0.09
Light jam	0.16	0.07	-0.10	-0.06	-0.04
Honey	0.08	-0.07	0.12	0.05	0.07
Other meats	-0.01	-0.01	0.05	0.01	-0.02
Other vegetables	0.17	-0.08	-0.03	0.01	-0.09
Other grains	0.03	-0.12	0.11	-0.07	0.02
White bread	-0.20	-0.07	0.18	-0.03	-0.05
Whole wheat bread	0.23	0.05	-0.13	0.02	0.02
Potatoes and Tubers	-0.09	-0.12	0.24	0.00	-0.20
Regular Pasta	-0.13	-0.10	-0.01	0.01	-0.11
Whole wheat pasta	0.03	0.01	-0.04	-0.05	0.01
Fresh fish	0.11	-0.10	0.06	0.01	-0.10
Canned fish and seafood	0.13	-0.11	0.01	0.03	-0.05
Light fresh cheese	0.13	0.06	-0.08	-0.04	-0.01
Cheese	0.10	-0.19	-0.03	0.10	-0.05
Sauces and dressings	-0.04	-0.24	-0.26	0.04	-0.09
Powder soups,...	0.02	0.01	0.08	-0.06	0.07
Tea	0.00	0.10	0.24	-0.13	-0.08
Wine	0.06	-0.03	0.07	0.08	-0.23
Vitamins and supplements	0.11	0.02	0.08	0.01	0.07
Regular yogurt and dairy desserts	0.04	-0.22	0.02	-0.04	0.31
Light yogurt and dairy desserts	0.23	0.07	-0.08	-0.06	-0.04

Appendix G: Food's average consumption per Cluster

Table G.1: Average intake frequency of foods aggregation level 2

	Clust. 1	Clust. 2	Clust. 3	Clust. 4	Clust. 5	Clust. 6
Sunflower oil	27.09	14.23	20.27	6.12	6.24	16.15
Olive oil	1.01	1.72	3.03	1.06	0.47	8.73
Edible oils	1.44	13.38	8.27	22.94	21.95	8.14
Mineral water	1.40	3.83	2.68	2.14	2.30	6.74
Liquors	0.29	2.95	0.29	0.21	0.14	0.62
White rice	8.01	10.76	8.53	8.65	8.32	7.12
Integral rice	0.06	0.02	0.08	0.03	0.02	0.20
Sugar	24.19	24.24	23.53	24.16	21.45	4.87
Regular soft drinks	9.23	15.96	9.60	14.95	7.43	2.95
Light soft drinks	0.82	1.31	0.57	0.59	0.77	5.46
Berries and Grapes	3.64	3.53	9.16	6.19	3.60	7.13
Coffee	8.03	14.62	12.44	5.78	4.21	10.54
Poultry	6.19	8.08	8.02	7.22	5.90	7.85
Pork	1.69	3.04	2.47	2.30	1.50	1.04
Lamb	0.22	0.36	0.96	0.12	0.10	0.15
Beef	6.05	9.63	8.18	7.60	4.28	6.28
Processed meats	6.87	12.91	9.44	11.67	5.19	8.07
Breakfast cereals	3.20	1.90	6.25	7.13	1.64	6.92
Beer	0.63	5.40	0.61	0.41	0.25	0.66
Chocolate	1.66	3.25	2.42	5.65	0.66	1.99
Citric fruits	5.43	5.33	11.69	6.75	4.00	9.72
Cream	0.14	0.67	0.35	0.35	0.06	0.32
Crustaceans and Molluscs	0.29	1.26	0.89	0.45	0.27	0.60
Soy derivatives	0.08	0.14	0.23	0.16	0.03	0.31
Sweeteners	4.23	4.23	5.80	2.26	7.44	23.41
Canned fruit	1.28	2.25	3.81	1.93	1.08	3.59
Fresh fruits	11.81	12.01	20.34	15.05	11.02	18.16
Dried fruits	0.40	0.53	1.97	0.73	0.29	1.01
Nuts	0.83	1.56	2.99	1.41	0.50	3.21
Light or wholeweat cookies	0.81	0.45	1.62	0.41	0.85	6.80
Wheat flour and other cereals	1.40	0.61	4.14	0.85	1.45	0.90
Ice cream and Ice	3.64	4.93	4.94	11.48	2.47	3.48
Fruit vegetables	15.08	16.78	23.33	18.63	16.63	21.99
Root vegetables	19.47	21.23	27.40	20.73	20.57	23.53
Leaf vegetables	14.68	15.64	23.70	16.77	15.10	21.42
Eggs	6.96	9.22	10.14	9.21	6.15	6.78
Juices	14.59	17.55	18.20	21.55	12.27	17.00
Whole milk	6.72	5.21	9.88	12.17	4.72	2.67
Skim milk	5.43	3.33	6.91	4.93	4.40	14.26
Other animal milk	0.02	0.03	0.01	0.01	0.01	0.00
Legumes	3.17	2.58	4.63	2.90	2.64	2.74
Manjar and cream derivatives	0.77	0.92	2.75	2.39	0.58	0.71
Animal butter	0.12	0.09	0.36	0.04	0.09	0.04
Butter and margarine	14.19	16.01	18.24	17.50	12.12	12.59

Table G.1: Average intake frequency of foods aggregation level 2
(continued)

	Clust. 1	Clust. 2	Clust. 3	Clust. 4	Clust. 5	Clust. 6
Sweet pastries	4.69	5.47	6.79	13.15	3.18	4.12
Jam	3.41	2.95	7.67	4.16	2.42	2.96
Light jam	0.12	0.26	0.25	0.13	0.25	3.16
Honey	0.47	0.29	2.86	0.57	0.38	0.94
Other meats	0.11	0.25	0.39	0.13	0.22	0.13
Other vegetables	0.43	1.26	1.47	0.54	0.30	2.00
Other grains	0.57	0.64	1.88	1.27	0.55	0.80
White bread	25.76	26.23	26.63	26.10	26.31	18.40
Whole wheat bread	0.80	0.52	0.89	0.58	0.49	8.09
Potatoes and Tubers	11.07	13.79	15.30	11.85	11.96	8.56
Regular Pasta	6.48	8.34	7.01	7.30	6.61	4.74
Whole wheat pasta	0.00	0.02	0.05	0.03	0.06	0.16
Fresh fish	1.23	1.80	2.28	1.45	1.08	2.01
Canned fish and seafood	1.83	2.49	3.51	2.29	1.52	3.33
Light fresh cheese	0.03	0.00	0.05	0.01	0.03	0.92
Cheese	6.86	10.28	11.01	8.68	4.31	9.30
Sauces and dressings	3.23	9.32	3.76	7.38	1.94	2.99
Powder soups, ...	0.06	0.00	0.36	0.06	0.36	0.16
Tea	21.72	16.69	22.51	18.14	25.89	22.11
Wine	0.88	2.61	1.90	0.24	0.80	1.89
Vitamins and supplements	2.25	0.48	3.87	1.37	2.09	5.33
Regular yogurt and dairy desserts	7.97	6.56	12.98	14.35	5.64	6.90
Light yogurt and dairy desserts	0.58	0.91	1.07	0.39	0.80	8.04