Analysis of Olive Oil Consumption in Japan Using Scanner Data

-Focusing on Health Consciousness of Consumers-

January 2020

Seifeddine Ben Taieb

Analysis of Olive Oil Consumption in Japan Using Scanner Data

-Focusing on Health Consciousness of Consumers-

A Dissertation Submitted to

the Graduate School of Life and Environmental Sciences,

the University of Tsukuba

in Partial Fulfilment of the Requirements

for the Degree of Doctor of Philosophy in Agricultural Science

(Doctoral Program in Appropriate Technology and Sciences for Sustainable Development)

Seifeddine Ben Taieb

Table of Contents

List of tables

List of Figures

Chapter 1 : Introduction

1. Overview

This thesis mainly discusses the effects of health consciousness on consumer behavior in Japan. The first part of the thesis focused on the Japanese oil market. We then discuss the consumption of olive oil in Japan.

Olive oil is mainly produced in the Mediterranean region. Spain, Italy, and Greece are the major producers followed by Tunisia, Morocco, and Turkey (according to FAOSTAT data). It is also in this region that olive oil is mainly consumed. Olive oil is considered a staple food in the Mediterranean diet, its health properties makes it a product with a promising future.

The consumption of olive oil has increased in recent years, giving rise to new emergent markets such as the United States of America, Canada, Japan, Brazil, China, Australia, etc. (according to FAOSTAT data). However, each market has its own unique characteristics and consumer base. Numerous studies regarding olive oil have been conducted in these markets.

Though the olive oil market in Japan has been growing at a phenomenal rate, not many studies have been conducted on olive oil consumption in Japan. Indeed, the few studies that were conducted focused on the price and country of origin of olive oil. Effects of health consciousness on olive oil consumption in Japan have never been studied.

2. Japanese oil market

Table 1-1 summarizes the recent situation of the Japanese oil market (from 2012 to 2016). In 2012, the market share of rapeseed oil was the highest followed by olive oil. However, by 2016, olive oil's market share jumped to the first place, surpassing rapeseed oil. This change was the result of sharp increase in olive oil's sales and fall in the sales of rapeseed oil. The other vegetable oils listed in the table (corn oil, palm oil, soy oil, and sunflower oil) did not show any significant change during this period. The other edible oils category however saw a spectacular growth during this period with the sales almost doubling in 5 years.

This change in market share amongst these vegetable oils is intriguing; ergo, we decided to focus on the Japanese oil market in the beginning of this thesis.

Table 1-1: Retail value sales of vegetable oil products in Japan (in US$ millions)

Categories	*2012*	*2013*	*2014*	*2015*	*2016*	*2012-2016 (compound annual growth rate %)*
Edible oils	**1,213.30**	**1,258.40**	**1,376.70**	**1,449.60**	**1,509.30**	**5.6**
Olive oil	264.3	301.1	328.1	351.5	373.2	9
Corn oil	63.1	62.7	61.2	60.3	59.6	-1.4
Palm oil	27.3	29.5	30.2	30.7	31.3	3.5
Rapeseed oil	405.5	377.2	365.9	354.2	339.1	-4.4
Soy oil	66.4	63	60.4	56.4	53.9	-5.1
Sunflower oil	23.2	23.2	23.3	26.2	28.5	5.3
Other edible oils	363.6	401.7	507.6	570.2	623.7	14.4

Source: Euromonitor (2007 cited in Alexander Perrault 2017, Vegetable Oil Product in Japan, Agriculture and Agri-food Canada (http://www.agr.gc.ca))

3. Japanese olive oil market

In the last two decades, consumption of olive oil in Japan has tremendously increased, showing an increasing interest of Japanese consumers in the Mediterranean diet. Olive oil market in Japan has been expanding due to dietary and health concerns.

The Japanese olive oil imports have been increasing gradually since 1996 (annual imports never reached 10,000 tons before 1996). (Mtimet et al., 2011)

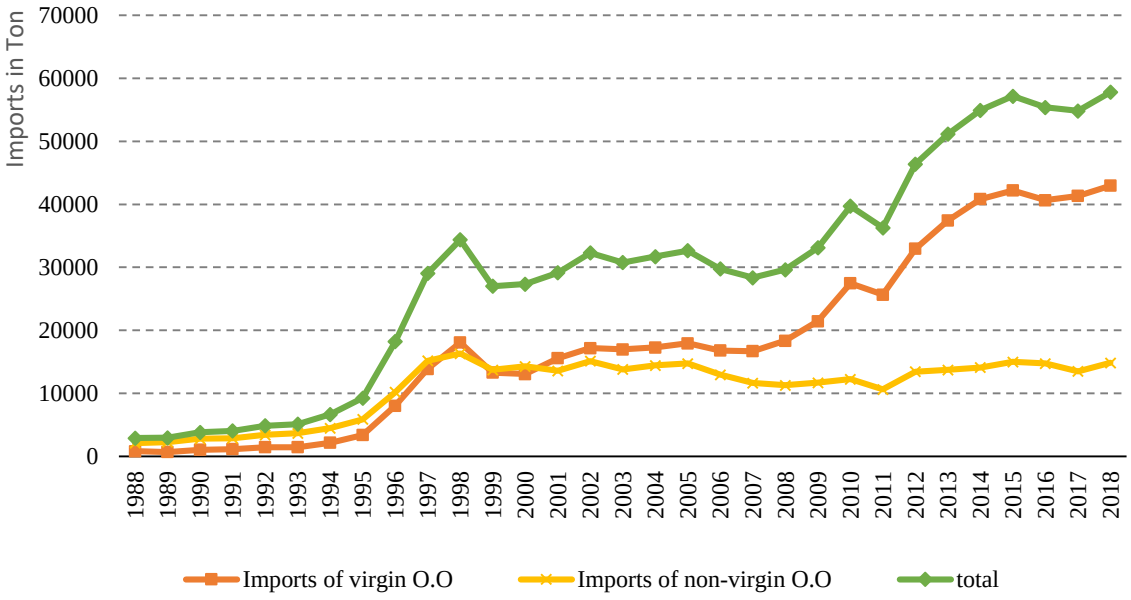
3.1. Japanese olive oil imports

Since Japan's olive oil production is negligible as compared to its olive oil imports, we can say that the imported quantities represent the total olive oil consumption in Japan.

Japan is considered as one of the major olive oil importers in the world. Japan's olive oil imports exceeded 30 thousand tons per year and reached 46 thousand tons in 2012 (Figure 1-1).

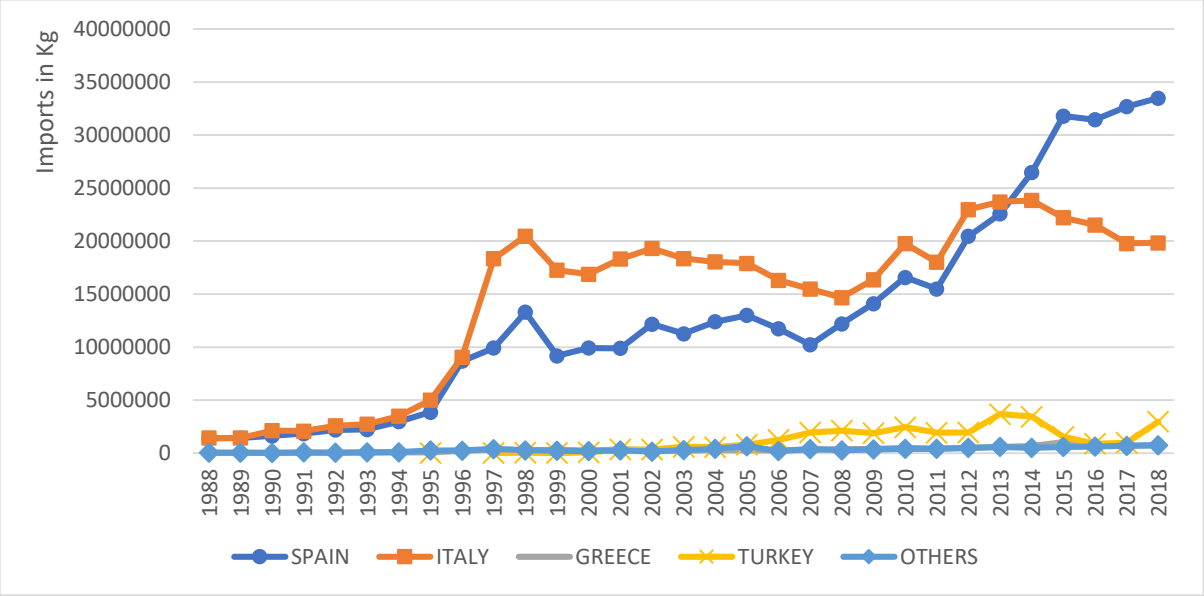
The trend of olive oil imports has also changed during these years. In fact, in the 80s and 90s, virgin olive oil imports were lower than non-virgin olive oil but since 2000 this situation has reversed. Virgin olive oil imports started increasing after 2000 and in 2012 the import quantity of virgin olive was twice that of non-virgin olive oil (Figure 1-1) which might imply that the Japanese consumer has become more aware of the quality of olive oil. According to Mtimet et al. (2008), the increase in consumption of olive oil in Japan is due to dietary and health concerns.

So far, Italy and Spain have sort of monopolized the Japanese olive oil market (figure 1-2). In 2012, the market share of Italy and Spain together was over 92%. Nevertheless, there are countries such as Turkey whose market share started increasing recently (figure 1-2). As for the other trading partners, their market share is practically insignificant.



Source: own elaboration from Japan trade statistics data

Figure 1-1: Japan's import of olive oil (by type)



Source: own elaboration from Japan trade statistics data

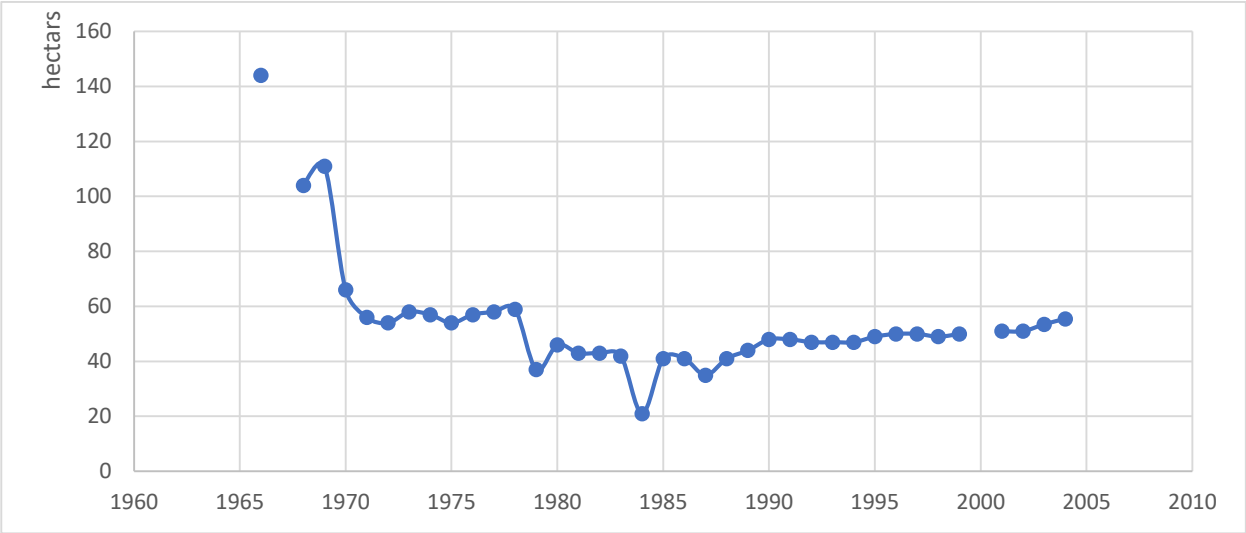
Figure 1-2: Japan's import of olive oil (by country)

3.2. Japanese olive oil production

Japan's olive production is limited to the southern parts of the country where the climate is similar to that of the Mediterranean region.

Compared to the 60s, the surface area dedicated to olive cultivation in Japan is quite smaller now. In fact, as we can see in figure 1-3, the surface area decreased from around 140 hectares to 60 hectares in 2004.

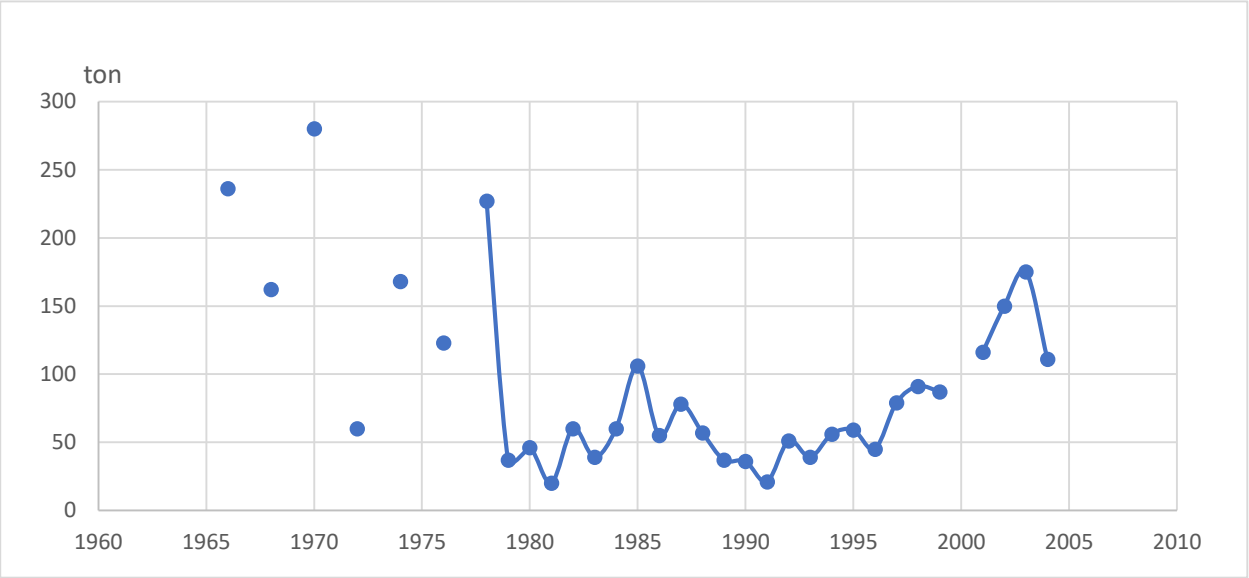
The olive production in Japan is low compared to other countries. Olive production is also relatively new to Japan; in fact, Japan started planting olive trees in 1908. The production of olives in Japan has increased recently; it reached 175 tons (figure 1-4) in 2003.



Source: own elaboration from the Japanese ministry of agriculture's data

Note) Missing numbers from the data

Figure 1-3: Area occupied by olive trees in Japan



Source: own elaboration from the Japanese ministry of agriculture's data

Note) Missing numbers from the data

Figure 1-4: Olive production in Japan

4. Health Consciousness

An increasing number of consumers now consider health issues to be one of the most important factors in their purchasing decisions. Manufacturers use labels to emphasize the health-related characteristics of their products in order to improve their sales (Nagata and Yamada, 2008). This has become a global trend. Indeed, in Northern Europe, consumers appear to follow a healthier dietary profile by increasing their consumption of fruits and vegetables, fish, and seafood, and reducing their fat intake (Kearney, 2010). Jussaume and Judson (1992) found that consumers in Kobe, Japan were significantly more concerned about food safety than residents of Seattle, USA.

This trend of consuming healthier products is one of the reasons to conduct the research presented in this thesis. As shown in table 1-1, the market share of Canola oil (a.k.a. Rapeseed oil) decreased in the 5-year period (2012-2016) while that of olive oil, which is considered healthy, increased tremendously.

5. Aim and structure of this thesis

This thesis aims to:

- Study the effects of health consciousness and other sociodemographic characteristics on oil purchase choice.
- Study the olive oil consumption patterns in Japan.
- Study the effects of health consciousness and other sociodemographic characteristics on the olive oil consumption.

Accordingly, this thesis is composed of 4 main parts, as shown in figure 1-5. These are: literature review, effects of health consciousness on oil consumption, olive oil consumption patterns in Japan, and effects of health consciousness on olive oil consumption.

- In the literature review chapter, we introduced papers that have either used the same models that we used in this thesis or studied olive oil consumption behavior.
- The third chapter focuses on the effects of health consciousness on oil consumption in Japan. We used the Dirichlet model to analyze the oil purchase data in Japan. These data were combined with survey data on the health aspects of consumers (health condition, lifestyle habits, etc.).

- The fourth chapter explores the olive oil consumption patterns in Japan. In fact, before studying the effects of health consciousness on olive oil consumption, we decided to study the Japanese olive oil market as an introductory research. In this chapter, the multinomial logit model was used to analyze purchase history data of olive oil.
- The fifth chapter introduces our research on the effects of health consciousness on olive oil consumption in Japan. To study this effect, we opted for the hierarchical Bayesian approach to analyze the scanner data and survey data, which are similar to the data in the third chapter.

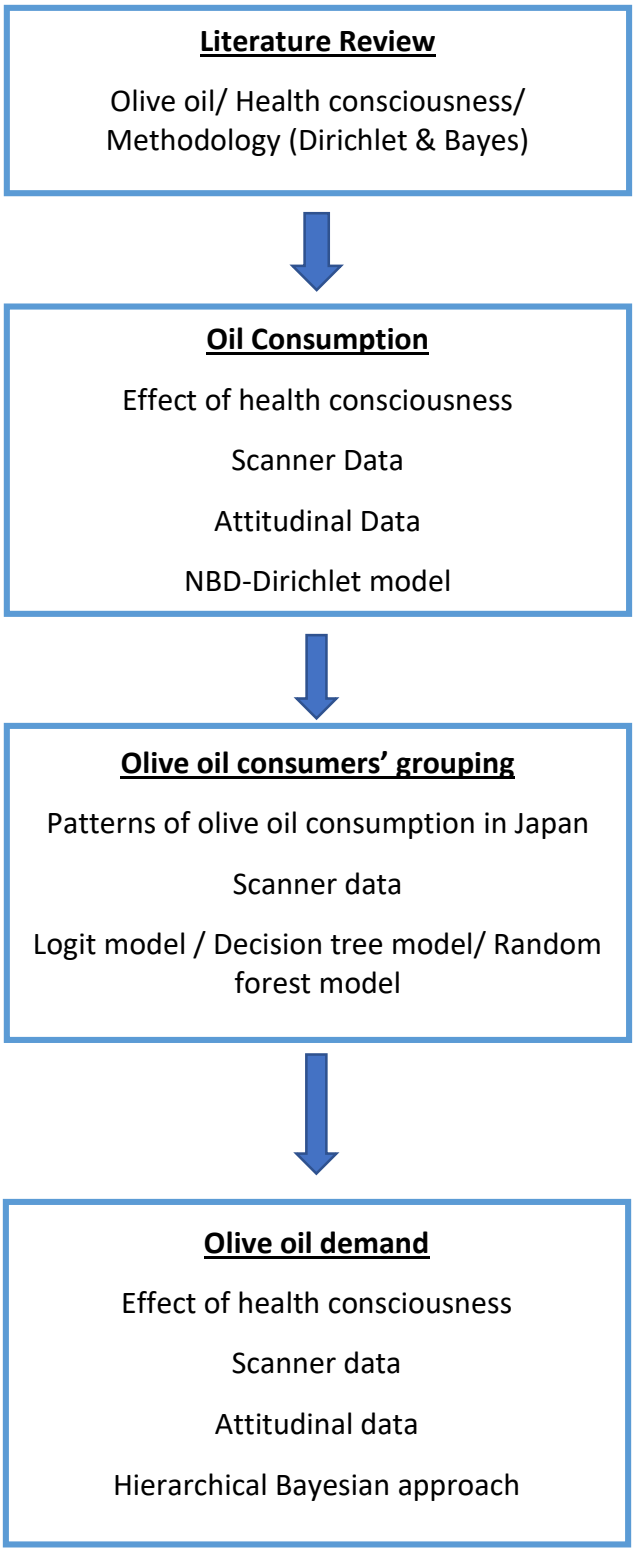


Figure 1-5: Structure of thesis

Chapter 2 : Literature review

1. Olive Oil Consumption

A study conducted by Yen and Chern (1992) found that unlike animal fat consumption, vegetable fat and oil consumption in the US has increased between 1950 and 1990. Japanese diet on the other hand, unlike the European and American diet, is mostly constituted of carbohydrates (rice). However, in the second half of the twentieth century, rice consumption per capita decreased with an increase in westernization of eating habits (Kim and Chern, 1999). One of the most important changes in Japanese diet was the increase in daily fat intake from 21 g per capita in 1955 to 58 g in 1992 (Sugano, 1996).

According to Delgado and Guinard (2010), the olive oil is an important component for most of the countries in the Mediterranean region (i.e., Spain, Italy, Greece, Tunisia, etc.), however, it is a relatively new product in areas outside of it. Inarehos-Garcia et al. (2010) has stated that "The fine flavor (aroma and taste) and color of virgin olive oil distinguish it from other edible vegetable oils, giving it a superior quality that is traditionally appreciated by consumers in Mediterranean countries and now all over the world.

Depending on the country, the consumers' perception of olive oil changes greatly. Santosa et al. (2010) found that in the USA, consumers were more knowledgeable about imported extra-virgin olive oil than the locally produced extra virgin olive oil with about two-thirds of them consuming the former. Caporal et al. (2006) have found that "in Italy, Information about origin affects the expectations with regards to specific sensory attributes in familiar consumers". Vlontzos and Duquenne (2013), have stated that the olive oil has been an essential part of the Greek diet. It has strong and long-life relations with Greek consumers, as it has always been used not only to fulfill nutritional needs but also for cultural and religious purposes.

In Japan, olive oil has now become common with imports increasing sharply since 1996.

1.1. Country of origin and geographical indication effect on Olive oil Consumption

Several papers have so far studied the impact of geographical origin on consumers' preferences. In the Mediterranean countries, consumers are concerned about the region of origin of olive oil since most of these consumers buy olive oil produced in their own countries. On the other hand, consumers from other countries view the country of origin as an important

attribute of the olive oil. Indeed, Dkhili et al. (2011) argued that in France, supermarkets carry olive oils from several countries; ergo, the French consumers place greater importance on the country of production than the region. Tunisians on the other hand place more importance on the region and olive variety since Tunisia does not import olive oil. Since they only consume Tunisian olive oils, Tunisians use these local differentiation criteria.

According to a study conducted in Albania, a country that produces olive oil, the majority of the respondents preferred higher priced domestic olive oil, while a small percentage preferred imported olive oil (Chan-Halbrendt et al., 2010). Aprille et al. (2012) found that consumers in Italy are willing to pay the highest premium price for a product with a PDO (Product Designation of Origin) label, followed by a product with an OF (organic Farming) label, and a PGI (Protected Geographical Indication) label. Examples of both Albania and Italy show that consumers from olive oil producing countries tend to prefer local products (olive oil produced in their own country) and are thus more concerned by the region of production rather than the country.

On the other hand, for consumers from countries outside the Mediterranean region where the olive oil market mainly depends on imports, the country of origin is one of the most important attributes. Mtimet et al. (2011) studied the effect of country of origin on Japanese consumers' olive oil preferences and found out that the latter consider Italian olive oil the best when making purchasing decisions.

Cavallo and Piqueras‐Fiszman (2017) studied the effect of olive oil attributes on the healthiness perception and found out that the most important element which influences the healthiness perception of extra virgin olive oil is its origin.

1.2. Quality of Olive oil

The perception of quality also seems to be different between producer countries (Mediterranean region) and countries that mainly import olive oil.

Servili and Montedoro, (2002) stated that "the quality of virgin olive oil is mainly connected to several specific activities of the hydrophilic phenols including their antioxidant power and other properties that affect the health and sensory aspects of virgin olive oil". Roselli et al. (2017) has argued that competition based on cost reduction strategies between producers has negative effects on profitability and consumers' difficulties in evaluating the quality of

olive oil is one of the reasons behind this. In their study, they established that health claims represent an unexploited tool that could be used to segment the olive oil trade.

Krystallis and Ness (2005) have used a conjoint analysis to analyze consumer preferences for "quality" olive oil in Greece. Their findings show 6 clusters of consumers with one of them labeled "highly health and quality conscious." This cluster showed very high awareness of quality schemes and attributed the highest importance to health information followed by ISO certification and an average importance to country of origin and HACCP (Hazard Analysis & Critical Control Point) certification.

Santosa et al. (2013) has established that as long as consumers perceive the olive oil to be of good quality, they think that they are getting a good value for whatever price they paid.

In Japan, we have found in a previous study that consumers rate olive oil quality based on country of origin, packaging, and taste. They have a high preference for Italian olive oil and fruity taste. Additionally, they prefer bottles with caps that prevent leakage.

2. Health consciousness and consumption

2.1. Health consciousness

Health consciousness reflects consumers' readiness to act to either improve their health or stay healthy. Health-conscious consumers are aware and concerned about their state of well-being and strive to make it better and/or keep their health and quality of life, as well as prevent ill health by engaging in healthy behaviors such as having a healthy diet and being self-conscious regarding health (Kraft and Goodell, 1993; Newsom et al., 2005; Mai and Hoffmann, 2012).

Chen (2009) has studied the effects of health consciousness on the consumption of organic products in Taiwan. The author used Likert scale questions to define health consciousness, healthy lifestyle, and environmental attitudes. The following questions were asked:

- Health consciousness: I have the impression that I sacrifice a lot for my heath/ I think I take health into account a lot in my life/ I often dwell on my health/ etc.

- Environmental attitudes: I prefer consuming recycled products/ I dispose of my garbage in different containers/ etc.
- Healthy lifestyle: I follow a low-salt diet/ I am a vegetarian/ I do exercise regularly/ etc.

Ellison et al. (2013) has studied among other things the effect of health consciousness on the caloric intake in restaurants. This study also employed 5-point Likert scale questions to assess the health consciousness of consumers. Questions on the following three topics were asked:

- Daily caloric intake, Fat intake, and Use of nutrition labels.

The authors then summed the answers to create a health consciousness variable with a value between 3 and 15.

In all the above studies, researchers used consumers' attitudes to reflect their health consciousness; however, other ways have also been employed to assess health consciousness. Mai and Hoffmann (2012) used Gould's (1988) method to assess health consciousness by asking Likert scale questions about how aware consumers are about their health. Gould himself had derived the method from the paper of Fenigstein et al. (1975) on self-consciousness.

Hong (2009) compiled many of the health-consciousness scales created by previous researchers (Gould, 1988; Kraft and Goodell, 1993; Tai and Tam, 1997; Jayanti and Burns, 1998; Michaelidou and Hassan, 2008; etc.) and combined them to re-conceptualize the health consciousness scale. Additionally, the author defined 5 dimensions of health consciousness:

- Integration of health behaviors: This dimension focuses on the behavior and attitude of consumers towards improving their health such as physical activity or having a healthy diet. It is important to note that this dimension was the focus of the present thesis.
- Psychological/ Inner state: This dimension posits that health consciousness is not visible through consumers' behavior but is related to the psychology of the consumer.

- Health information seeking and usage: As the name suggests, this dimension refers to how knowledgeable consumers are about health information. Ergo, health consciousness would be proportional to their knowledge of health information.
- Personal responsibility: This dimension posits that health consciousness reflects the responsibility towards oneself to manage personal health.
- Health motivation: The fifth and last dimension reflects the motivation of consumers to engage in preventive health care activities.

Hong (2009) concluded that health consciousness englobes 3 concepts: self-health awareness, personal responsibility, and health motivation.

2.2. Olive oil and health

Several papers have studied and reported the health benefits of extra virgin olive oil. It has been proven that some components of olive oil associated with these health effects are monounsaturated fatty acids (Martinez-Gonzales and Sanchez-Villegas, 2004; Tuck and Hayball, 2002; Covas et al., 2006) and polyphenols (Saija and Uccella, 2001; Gorzynik-Debicka et al., 2018). The olive oil also has the highest amount of squalene among vegetable oils (Owen et al., 2000). Covas et al. (2006) has established that olive oil-rich diets can be a useful tool against risk factors for cardiovascular disease.

The olive oil components listed above are found only in extra virgin olive oil due to the extraction process involved in it.

3. NBD-Dirichlet Model

The NBD-Dirichlet model is a combination of 2 models, the NBD model (i.e., Negative Binomial Distribution) and Dirichlet model (Dirichlet Multinomial Distribution (DMD)).

It was developed by Goodhardt, Ehrenberg, and Chatfield (Ehrenberg 1959; Goodhardt, Ehrenberg et al., 1984). This model is special, among other reasons, because of its predictive power which lies in the shape of the distributions. It has been mostly used to analyze brand loyalty among different products from the same category. The changes in brand market shares in a repertoire market are closely connected to some sort of repeat purchasing behavior. It is then safe to assume that this repeat purchase behavior is connected to certain degrees of brand loyalty, which may vary across categories (Jarvis et al., 2003).

Dirichlet Modeling allows us to generate estimates of the brand performance measures (BPMs) such as purchase rate, market share, penetration, purchase frequency, share of category requirements, and 100% loyal consumers (Rungie, 2003). Rungie et al. (2013) used this model to analyze the brand loyalty towards detergent in France while Bassi (2011) analyzed the brand loyalty towards beer in Italy.

Though the NBD-Dirichlet model was mainly created to study brand loyalty, several studies have used this model for different purposes. Wrigley and Dunn, (1984a, 1984b, 1984c) found that "the NBD models may usefully and successfully be transferred from their original context of brand purchasing to the analysis of purchasing patterns at individual stores in a single city". Lam and Mizerski (2009) have used the NBD-Dirichlet model to study gambling patterns in Australia. To do so, the researchers used aggregated data on annual game playing frequency. Casini et al. (2009) used this model to study Italian consumers' loyalty towards wine attributes.

4. Bayesian estimation

Accoring to Gelman et al., (1996), iterative simulation methods have recently become popular tools in statistical analysis, especially in the calculation of posterior distributions arising in Bayesian inference. The Bayesian approach has been around for quite a long time. In fact, Maritz and Lwin (2018) trace the early examples of the use of Empirical Bayes data back to the 1940s. The Bayesian approach to modeling and data analysis is becoming increasingly popular as it is being seen as an effective and practical alternative to the frequentist approach. Indeed, due to advances in the computing field which enabled faster and better Bayesian designs and analyses, "the philosophical battles between frequentists and Bayesians that were once common at professional statistical meetings are being replaced by a single, more eclectic approach" (Carlin and Louis, 2010). Natarajan and Kass (2000) noted that the two-stage hierarchical models have made one of the major contributions to the Bayesian approach to data analysis.

The process of Bayesian statistical analysis requires incorporation of prior belief (information we have beforehand) in the model. In doing so, the Bayesian approach offers

solutions to several problems, such as how to analyze multiple exposures (Gelman et al., 2013). In fact, specifying a prior distribution is of utmost importance in Bayesian analysis. Expressing prior belief in the form of a distribution is often difficult, especially in multiparameter models (Ibrahim and Laud, 1991). The type of priors used in Bayesian estimations have changed through time. Jeffreys (1946) has defined rules and conditions for priors which later came to be known as Jeffreys prior. This prior has been used in several studies (for instance: Ibrahim and Laud, 1991). Jefferys' prior was also extensively discussed by Kass and Wasserman (1996a, 1996b), who have reviewed in detail the rules and process of selection of prior distributions. It was also discussed by Gelman (2006) in his paper on prior distributions for variance parameters in hierarchical models.

Natarjan and Kass (2000) noted 2 types of prior, a prior analogous to the conventional $1/\sigma$ prior in the one-sample normal (μ, σ^2) and diffuse conjugate prior. Both these priors are however not perfect; indeed the first prior does not lead to a proper posterior (i.e., does not integrate to a finite number; an unnormalized density $f(\theta)$ is proper if $\int f(\theta)d\theta=n$, with n being a finite number). On the other hand, the diffuse conjugate priors (Natarajan and McCulloch, 1998), which are priors having the same probability distribution as the posterior distribution, may lead to inaccurate posterior estimates. The benefit of using a conjugate prior is the ease of calculation; however, with the advancements in computation techniques such as MCMC, that benefit has been rendered moot.

In their book on Bayesian data analysis, Gelman et al. (2013) have used several distributions (Normal, inverse-χ^2, Student's t-distribution, etc.) for the priors in the illustrated examples. Additionally, Young and Berger (1996) have made a catalog of noninformative priors.

The Bayesian approach has been widely used not only in analyzing consumption behavior but also in other fields such as medicine. Other instances are as follows:

- Allenby et al. (1998) has used the Bayesian approach to study the demand heterogeneity. The authors used 3 different data sets; the first from a conjoint study of consumer preferences for marine engines and the second and third datasets are scanner data of food products.
- Yang et al. (2003) has used the Bayesian approach to estimate a heterogeneous demand and supply model, of light beer, with household panel data.

- Using the Monte Carlo Makrov Chain (MCMC) technique within the Bayesian framework, Wang et al. (2007) have estimated a multivariate Poisson regression model to predict the pattern of cross-category store brand purchasing behavior.
- Chandukala et al. (2011) has used the Bayesian estimation approach to investigate the purchasing behavior towards luxury cars.
- Byun (2017) has applied a discrete choice experiment and then used a hierarchical Bayesian logit model to analyze consumers' preferences for electricity generation source.
- Volinski et al. (2009) has applied a choice experiment and then used a hierarchical Bayesian approach to study the behavior of Canadian consumers towards GM (genetically modified) labels on Canola oil.
- Kim and Sugai (2008) have also used a hierarchical Bayesian approach to analyze data gathered from a survey on internet-based services in Japan. The objective of this study was to determine the willingness of consumers to pay for those services.
- Kasteridis and Yen (2012) have used a Bayesian approach with a Tobit base to study the demand for organic vegetables in the USA. The data used in this paper were the Nielson homescan panel data.

5. Thesis originality

The methodological originality of this thesis mainly centers around the data and the way the models were used. First, the data used in this thesis were scanner data which up to now haven't been used in Japan to analyze consumer behavior related to olive oil. Moreover, the scanner data were combined with attitudinal data on the health aspects of consumers. The use of two types of data, scanner data and survey data, constitutes one of the main originalities of this thesis. Second, the Dirichlet model was used in a new way in this thesis to analyze consumer purchase choice. So far, this model was mainly used to analyze brand loyalty, and, in a few instances, store loyalty. Using it this way, to our knowledge, was a first.

From an empirical viewpoint, effect of health consciousness on olive oil consumption in Japan hasn't been studied before. Even though olive oil has one of the largest market shares in Japan, not many studies have been conducted on its consumers.

Although, there are studies on the effect of health consciousness on oil consumption in Japan, none of them have employed scanner data to do so.

Chapter 3：Effect of Health Consciousness on Oil Consumption in Japan

1. Introduction

A recent report on vegetable oil products in Japan (Perrault, 2017), which analyzed a four-year period (2012–2016), stated that the edible oil market has grown by 5.6%. Accordingly, its retail value grew from US$1,213.3 million to US$1,509.3 million. In 2016, olive oil jumped to first place in terms of retail sales in Japan, with a value of US$373.2 million in 2016, putting rapeseed oil (also known as canola oil) in second place in Japan, with a retail value of US$339.1 million. In fact, sales of rapeseed oil and soy oil have decreased in these four years by 4.4% and 5.1%, respectively. The sharp growth in demand of "other edible oils" in Japan suggests that Japanese consumers are seeking new type of oils. Hence, this study aimed to investigate such oils' demand and to understand how health consciousness affects the demand for each type of oil in the Japanese market.

The Japanese consumer is probably becoming more health-conscious and thus seeking healthier alternatives such as perilla, linseed, or olive oil, the last of which is one of the most consumed vegetable oils in Japan (Ben Taieb and Ujiie, 2018). Kraft and Goodell (1993) have noted that individuals who have a healthy lifestyle are concerned with their diet, physical health, stress, and the environment. In Japan, Takeshita (1999) analyzed newspaper articles and found that the increase in demand of premium oil was due to increased interest in learning about health risks. Because different fats and oils have different saturated fat contents, each may be affected differently by health risk information (Chern et al., 1995), and based on that information, consumers might assign products to different levels of healthiness. For instance, among seed oils, flaxseed oil had the highest percentage (57%) of omega-3 fatty acid, a-linolenic acid (Hasler, 1998; Gebauer et al. 2006; Tonon et al. 2011).The olive oil on the other hand contains a considerable amount of Squalene (a highly unsaturated aliphatic hydrocarbon with important biological properties) which significantly contributes to its health claims (Gunstone, 2002).

This chapter shows how health consciousness affects oil consumption. In other words, the chapter discusses the oils health-conscious consumers choose.

2. Data

2.1. Scanner panel data

Consumer behavior was analyzed using this data. Scanner data are a collection of the purchase history of certain products. For this study, the data were provided by a Japanese marketing company, which recruited consumers (referred to as monitors) and gave each of them a scanner. Whenever monitors purchase a product, they would scan its bar code, and all information about that product, including date and time, would be automatically added to the database. The scanner panel data were obtained from Macromille Inc., a Japanese marketing company. These data cover purchase history of oil for 13,262 households in Japan for a 2-year period (2015–2016). After data cleaning, only those consumers were retained who were part of the monitoring process for the entire 2 years, who purchased an oil product at least once during these 2 years, and who answered the questionnaire.

Scanner data constitute a source of product-specific information (Nayga and Oral, 1994). In fact, they show revealed preferences of consumers—the actual choices of consumers (Casini et al., 2009).

Using the scanner data, the market share of each type of oil was calculated based on the number of items bought for each type. According to these data, canola oil has the highest market share followed by sesame and olive oil[1] (Figure 3-1).

2.2. Survey attitudinal data

From the survey data, we selected a few attitudinal data to be used as variables in the model. These attitudinal data represent different types of health consciousness among consumers. Moreover, we selected a few health problems that might be affected by oil consumption, such as blood pressure and cholesterol. Alonso and Martínez‐González (2004) noted that in some small studies and clinical trials, olive oil has been shown to lower blood pressure. Table 3-2 lists the variable used in the model as well as their descriptive statistics. In this study, we divide health consciousness into three categories by calculating the mean of many variables obtained through Yes (1) / No (0) questions (Table 3-1): sports activities, healthy lifestyle, and

[1] The olive oil is a very differentiated product, its price and quality vary a lot. However, in this chapter we didn't take that into account.

healthy diet. We used these attitudinal and behavioral variables as proxy variables that capture the health-conscious mindset of consumers. Since the food variable shows a general inclination toward a healthy diet and is not specific to oil consumption, it was assumed that it is independent of the oil consumption error.

Finally, the health problems category depicts illnesses that are related to fat consumption. Though certain health conditions can motivate consumers to be more health-conscious, it is not a definite equivalency, and thus, it is assumed that having health problems does not automatically translate to healthy food habits.

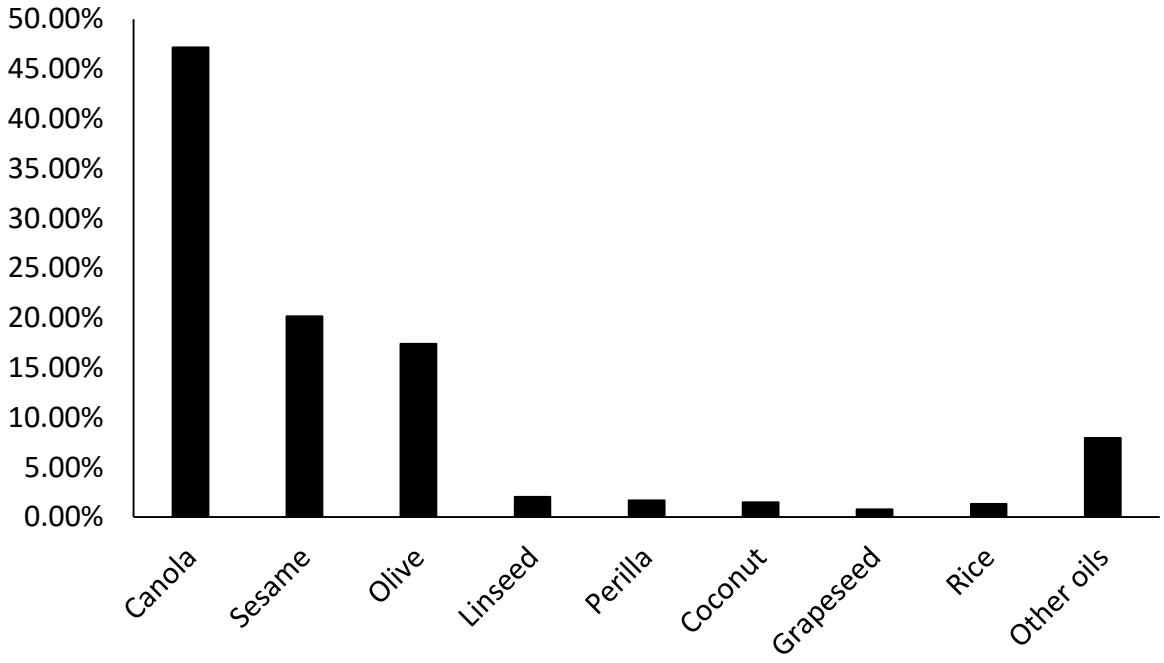


Figure 3-1: Oil Market Share (Source: Compiled by the authors)

Table 3-1: Questions used to reflect health consciousness of consumers

Please select what you do for your health	
Walking	
Jogging	
Gymnastics / stretching	
Muscle training	
Other sports	
Keep track of bodyweight and body fat	
Keep track of blood pressure	
Maintain a regular schedule	
Have enough sleep	
Regularly gargle and wash hands	
Relieve stress	
Have massages	
Have a balanced diet	
Reduce intake of salt	
Reduce intake of sugar	
Reduce intake of fat	
Keep track of calorie intake	
Make food at home instead of eating out	
Eat a lot of vegetables	
Eat meat and fish	
Other	

Source: Survey data

Table 3-2: Explanatory Variables

Category	Variable	Interval	Mean	Variance
Demographic Variables	Age	16 ~ 76	49.9	180.05
	Household income (1)	1 ~ 14	5.83	14.96
	Household size	1 ~ 10	2.67	1.55
Health consciousness variables	Sports	0 ~ 1	0.14	0.03
	Lifestyle	0 ~ 1	0.25	0.05
	Food	0 ~ 1	0.26	0.27
Health problems	Diabetes	0 or 1	0.12	0.11
	Blood pressure	0 or 1	0.15	0.13
	Cholesterol	0 or 1	0.13	0.12

Note (1): Household income: Minimum: 1 (2 million yen)/ Maximum: 14 (over 20 million yen)

Source: Compiled by the author

3. NBD-Dirichlet model

The NBD-Dirichlet model describes how frequently bought branded consumer products, such as instant coffee or toothpaste, are purchased when the market is stationary and unsegmented. It was developed by Goodhardt, Ehrenberg and Chatfield (Ehrenberg 1959; Goodhardt, Ehrenberg et al. 1984). This model has been widely used to analyze brand loyalty among different products from the same category, for instance, the beer market in Italy (Bassi, 2011) and laundry detergents in France (Rungie et al., 2013).

Other studies have used the NBD-Dirichlet model to analyze the store choice, and not the brand choice. Indeed, Wrigley and Dunn (1984a; 1984b; 1984c), have used the NBD-Dirichlet model to analyze purchasing patterns at individual stores in a single city.

In our study, however, the Dirichlet model was used not to assess brand choice or store choice but to assess the choice of oil (canola, olive, linseed, etc.) in the Japanese edible oil market.

Oil consumption has an NBD, while the choice of "type of oil" has a DMD. Both NBD and DMD are used to count data as dependent variable, which in our case is the number of items bought.

Equation (1) is the NBD with two parameters, which are both positive: the shape parameter *γ* and the scale parameter *β*,

$$f_{\gamma,\beta}(k) = \frac{\Gamma(\gamma+k)}{\Gamma(\gamma)k!}\frac{\beta^k}{(1+\beta)^{(\gamma+k)}} \quad (1)$$

where *k* is the number of items purchased and $\Gamma(\cdot)$ is gamma function.

The purpose of using the NBD is to calculate the expectation as in equation (2), of the number of items purchased, *k*, bought.

$$E(k) = \beta\gamma \quad (2)$$

The DMD model assumes that the purchases of each type are conditional on the category purchase rate. The Dirichlet multinomial distribution has h parameters, as seen in equation (3), one for each type of oil. These are α_1, α_2,..., α_h where each is positive.

$$f_{\alpha_1,\alpha_2,\ldots,\alpha_h}(r_1, r_2, \ldots, r_h|k) = \frac{\Gamma\left(\sum_{j=1}^{h}\alpha_j\right)k!}{\Gamma(\sum_{j=1}^{h}\alpha_j+k)}\prod_{j=1}^{h}\frac{\Gamma(\alpha_j+r_j)}{r_j!\Gamma(\alpha_j)} \qquad (3)$$

Where r_j is the number of items bought for oil type *j*.

In this study, we are interested in the effect of consumers' characteristics (age, household size, etc.) as well as health consciousness on the choice of oil. Consumers' characteristics are introduced in the Dirichlet model as covariates. The parameters of the Dirichlet model become functions of the covariates. The introduction of consumer characteristics in the Dirichlet model as covariates has been discussed at length by Wrigley and Dunn (1985) and Rungie et al (2013).

To introduce consumer characteristics, the parameters of the two models, NBD and DMD, are transformed as follows:

$$E(k_i|x_i) = \gamma\beta = e^{\delta' x_i} \qquad (4)$$

$$\alpha_j(x_i) = e^{\theta_j' x_i} \qquad (5)$$

The coefficients of the resulting DMD model affect the probability of purchase (P_{ij}) for each type "*j*" of oil by individual "*i*" as follows (Goodhardt et al., 1984):

$$P_{ij} = \frac{\alpha_j(x_i)}{\sum_{m=1}^{h}\alpha_m(x_i)} = \frac{e^{\theta_j' x_i}}{\sum_{m=1}^{h} e^{\theta_m' x_i}} \qquad (6)$$

where x_i is a vector of consumer i's characteristics, δ is a vector of the model's parameter for the category (total edible oils) and θ_j, for j= 1…h, the vectors of the model's parameters for type j. The δ and are θs$_j$ estimated using the maximum likelihood estimation method.

- In the NBD model, the dependent variable *K* denotes the total number of items of oil bought by the monitors.
- In the case of the DMD, the dependent variable k_j was calculated by summing the number of items bought by monitors. For each consumer *i* we had *h K*, (*h* being the number of oil types).

Table 3-3: Results of the Dirichlet model

	NBD (δs)	Dirichlet Multinomial Model (DMD) (θs)								
	Edible oil	Olive oil	Sesame oil	Canola	Coconut	Linseed	Perilla	Grape Seed	Rice	Other oils
intercept	0.77***	0.0634	0.5974***	0.3057**	-3.7126***	-3.6300***	-4.0779***	-3.2345***	-3.9399***	-1.1936***
age	0.017***	-0.0053***	-0.0160***	-0.0061***	0.0151***	0.0168***	0.0225***	-0.0003	0.0105*	-0.0030#
Hh_inc	-0.001	0.006	-0.0005	-0.0169***	0.0250*	0.0327***	0.0253*	0.019	0.0149	0.0038
Hh_size	0.241***	-0.0881***	0.0695***	0.2676***	-0.1366***	-0.1431***	-0.1260**	-0.1753***	-0.0571	0.0332#
Sports	-0.075#	0.2462*	0.0473	-0.3503***	0.8264***	0.3069***	0.6610**	0.0703	-0.3432	0.2391#
Lifestyle	0.141***	0.1762#	0.0979	0.0681	-0.1811	0.3637*	0.0648	0.4936#	0.3265	0.1209
Food	0.206***	0.1873*	-0.0005	-0.4497***	0.3942*	0.4067**	0.2970#	0.0782	0.7615***	-0.0411
Diabetes	-0.02	-0.1411**	-0.1156*	0.0428	-0.2533*	-0.1612#	-0.2922*	-0.1768	-0.3048*	-0.0366
Blood pressure	0.005					0.1723#	0.2790**	0.1342	0.1435	0.1453*
Cholesterol	-0.038#					0.3303**	0.2765*	0.2721#	0.0882	0.0536
Log Likelihood	-88029.335					-77726.19				

Significance code: ***, **, *, and # indicate significance at the 0.0001, 0.001, 0.01, and 0.05 levels, respectively.

Source: Compiled by the author

Table 3-4: Results of the Dirichlet model without the "food" variable

	NBD (δs)	Dirichlet Multinomial Model (DMD) (θs)								
	Edible oil	Olive oil	Sesame oil	Canola	Coconuts	Linseed	Perilla	Grape Seed	Rice	Other oils
intercept	0.7638***	0.0675	0.6080***	0.3359***	-3.7233***	-3.6404***	-4.0839***	-3.2311***	-3.9591***	-1.1838***
age	0.0176***	-0.0050**	-0.0162***	-0.0076***	0.0162***	0.0179***	0.0233***	-0.0002	0.0128**	-0.0033
Hh_inc	-0.0007	0.0061	-0.0006	-0.0174***	0.0255*	0.0331***	0.0256*	0.019	0.016	0.0037
Hh_size	0.2401***	-0.0907***	0.0672***	0.2660***	-0.1392***	-0.1459***	-0.1286***	-0.1771***	-0.0623#	0.0314#
Sports	-0.0325	0.2836**	0.0456	-0.4462***	0.9048***	0.3915*	0.7185***	0.0856	-0.1834	0.2290#
Lifestyle	0.2543***	0.2857***	0.1042	-0.1567#	0.0447	0.5943***	0.2356	0.5421*	0.7588***	0.1048
Diabetes	-0.0216	-0.1373*	-0.1100*	0.0571	-0.2551*	-0.1630#	-0.2928*	-0.1741	-0.3153*	-0.0314
Blood pressure	0.0113	0.0889#	0.0414	0.0054	0.1048	0.1856*	0.2887**	0.1371	0.1686#	0.1444*
Cholesterol	-0.0359#	0.0889#	-0.0043	-0.0658	0.4316***	0.3339**	0.2771*	0.2690#	0.0954	0.0483
Log Likelihood	-88071					-77821.23				
Significane code : '***' 0.0001, '**' 0.001, '*' 0.01, '#' 0.05										

Source: Compiled by the author

Table 3-5: *Results of the Dirichlet model with two types of oil*

	Dirichlet Multinomial Model (DMD) (θs)		
	Olive oil	canola	others
(Intercept)	0.0817	0.3462**	0.8003***
age	-0.0086***	-0.0101***	-0.0140***
hhinc	0.0068	-0.0158**	0.0052
hh_size	-0.0811***	0.2667***	0.0337*
sports	0.2587*	-0.3149**	0.1750#
lifestyle	0.1635#	0.0564	0.1088
food	0.1671*	-0.4597***	0.0327
diabetes	-0.1009#	0.0844#	-0.0714
blood_pressure	0.0798#	0.0146	0.0874#
cholesterol	0.0734	-0.0669	0.0361

Source: Compiled by the author

4. Results

The estimation results in Table 3-3 show negative binomial model results (first column) as well as Dirichlet multinomial results. The independent variables were in three groups: demographic characteristics (age, household income, household size), psychographic variables pertaining to health consciousness (sport, lifestyle, food), and chronic diseases (diabetes, blood pressure, cholesterol).

Table 3-4 and 3-5 represent alternative models to gauge the robustness of the original model. In Table 3-4, even after omitting the "food" variable, the results for the remaining independent variables do not change. The same could be said for Table 3-5. In this model, only two types of oils were used, "olive oil" and "Canola oil," and the rest were aggregated under "other oils." The results of this estimation show no significant difference between this model and the original one (Table 3-3). Hence, the original estimation result was adopted.

Edible oil consumption in Japan is positively affected by age and household size. The health consciousness psychographic variables, however, affect oil consumption differently. While consumers who practice sports buy less oil, those who lead a healthy lifestyle and those who have a healthy diet buy more oil. This could mean that those who practice sports would reduce their oil intake while others would probably consume more healthy oils, resulting in an increase in the total consumption of oil.

Finally, oil consumption seems to be only affected by cholesterol in the health status section. In fact, consumers with cholesterol-related health issues buy less oil products than those without.

The results of the Dirichlet Multinomial model are documented from the third column onward in Table 3-3.

Unlike in the total oil consumption results (NBD model) in the DMD model, age and household size seem to have an opposite effect. In fact, whenever the age coefficient is positive, the household coefficient is negative and vice-versa. Household income has negatively affected the purchase of canola oil while positively affecting coconuts, linseed, and perilla oil, all healthy oils higher-priced than canola oil. This suggests that households with a

high income would reduce their consumption of canola oil and increase their consumption of healthy oils such as linseed (flaxseed) or perilla.

The psychographic variables corroborate findings of previous research on the benefits of coconut oil (Pehowich et al., 2000; Carandang, 2008; DebMandal and Shyamapada, 2011), linseed oil (Popa et al., 2012; Tripathi et al., 2013), perilla seed oil (Kurowska et al., 2003; Adhikari et al., 2006), and rice bran oil (Nagendra Prasad et al., 2011). Indeed, health-conscious consumers have a higher purchase rate of coconut, linseed, perilla, and rice bran oil than that of canola or sesame oil. The psychographic variables pertaining to health consciousness confirm that consumers who wish to maintain good health consume more healthy oils. All three types of consumers—those who practice sports, those who lead a healthy lifestyle, and those who follow a healthy diet—have a high purchase rate of olive oil. On the other hand, the purchase rate of canola oil is lower for the first (sports) and third (healthy diet; "food") categories of consumers. Similar results can be seen for coconut, linseed, and perilla oils, where at least one category of health-conscious consumers have a high purchase rate of these oils. The difference in coefficients between the types of oil may reflect different uses of these oils. For instance, coconut oil has the highest coefficient among all types of oil for the "sports" variable, meaning that consumers who practice sports have the highest purchase rate of coconut oil, while having the lowest consumption of rice bran oil. On the other hand, the second category of consumers (lifestyle) have the highest purchase rate of grape seed oil. The third category of consumers (food) have the highest consumption of rice bran oil.

Finally, the last section of Table 3-3 illustrates how health problems affect the choice of oil. While consumers who suffer from diabetes have predominantly a smaller consumption of all types of oil—even the ones viewed as healthy—there is still a difference among the coefficients. Indeed, sesame oil, followed by olive oil, has a higher purchase rate among such consumers. Those with high blood pressure and cholesterol have a high purchase rate of the healthy oils. Consumers with hypertension have the highest purchase rate of perilla seed oil, while those with high cholesterol have the highest purchase rate of coconut oil.

5. Conclusion

In this study, our objective was to investigate the effects of health consciousness on the choice of oil among Japanese consumers.

Though oil consumption has increased in Japan, every type of oil has not seen a growth in consumption. In fact, canola oil consumption has decreased during the last 2 years, while consumption of healthy oils, such as olive oil, has increased.

The estimation results show that health-conscious consumers generally buy healthy oils (i.e., olive, linseed, etc.). However, not all categories of health-conscious consumers choose the same oil. Depending on how they are classified as health-conscious (practicing sports, leading a healthy lifestyle, or following a healthy diet), their top choice of oil was different.

The effect of health status on consumption was also different depending on the chronic disease. Indeed, while diabetic consumers have a smaller purchase rate in all the oil types, consumers with cholesterol and hypertension have a higher consumption of healthy oils than those without. Another interesting result was that consumers with high levels of cholesterol have the highest consumption of coconut oil. Though coconut oil was proven to be healthier than other oils, it could contribute to higher levels of cholesterol. This result indicates that consumers are not very knowledgeable about specific health effects of oils but only have a general idea about the health benefits of each oil.

This study analyzed 2 years of scanner data. Future research should analyze a bigger database (spanning more than 2 years), which would shed more light into this subject. Moreover, as only an empirical analysis was conducted in this study, a theoretical model to describe the relationship between health consciousness and oil consumption in Japan should be developed. Additionally, although we assumed that the oil consumption error is independent from the health consciousness variables, the endogeneity problem must not be ignored and remains a task for future studies. Finally, since count data were used instead of quantity, a price variable could not be included in the model. This model describes the effects of health consciousness on the purchase frequency of oils and not on the actual quantity of oil bought.

Chapter 4：Analysis of Olive Oil Consumers in Japan Using Scanner Data
-Focusing on Purchase Price and Quantity-

1. Introduction

Menapace et al., (2011) has stated that the olive oil has historically been significant in the Mediterranean countries. However, as this diet gained popularity worldwide, olive oil consumption grew considerably in many countries, including Australia, Brazil, Canada, Japan, and the United States". In fact, the Japanese olive oil market is one of the largest markets in the world today. Indeed, in 2016, olive oil imports reached nearly 60,000 tons (Figure 1-1, Chapter 1). Prior to 2000, olive oil imports mostly included non-virgin olive oil; however, after 2000, this trend reversed. In fact, virgin olive oil imports have surpassed non-virgin oil imports.

In 2016, retail sales of olive oil (at US$373.2 million) surpassed that of all edible oils in Japan, followed by sales of rapeseed oil (US$339 million). This value is expected to grow at a rate of 5.8% through 2021 (Euromonitor International, 2017). Olive oil is already one of the most widely used vegetable oils by Japanese consumers. Even though the product has occupied an important place within the Japanese market, we still do not have a good understanding of olive oil consumers. All we know so far is that the Japanese consumer prefers Italian olive oil and that they prefer the fruity taste (Mtimet et al., 2011).

In this study, scanner data were used to analyze consumer behavior related to olive oil. "In the 1980s, the emergence of scanner panel data constituted a major milestone for consumer-packaged goods manufacturers, retailers, and marketing scholars because the data provide deep insights into longitudinal consumer behavior" (Swait and Rick, 2003). To date, scanner data have not been used to analyze olive oil consumer behavior in Japan, although this approach has been used in other countries such as Spain, Italy, and the US.

In Spain, scanner data (obtained from a retail store) were used to study brand choice and investigate how a brand can gain popularity among consumers (Juan and Manuel, 2009).

In the US, a study using the Almost Ideal Demand System (AIDS) model was used to understand the difference in consumption behavior between consumers belonging to two different income groups (Jones et al., 2003).

In Italy, the AIDS model was also used to analyze scanner data. This study aimed to examine the relationships among the extra-virgin olive oil demand in the retailing sector, price vectors, and total purchases (Marchini et al., 2010).

So far, studies on consumer behavior have primarily used traditional models based on economic theories, such as the AIDS model and discrete choice models. However, with the advancement of models and emergence of new techniques such as data mining, consumer behavior research has experienced an expansion in methodologies. For example, Adriana (2013) has used the decision tree model, which is a data mining technique, to analyze the effect of knowledge and health consciousness on the consumption behavior of olive oil in Uruguay. The random forest model has been used by Plonsky (2017) to analyze human behavior. Plonsky et al., (2017), found that given any possible set of goods, the random forest model outperforms all other models. They argued that it was probably due to the fact that the stochastic and dichotomous nature of random decision trees align well with basic aspects of human decision making.

This brings us to one of the objectives of this study—to compare the traditional models with those from the data mining field.

The other objective of this study is purely empirical. Today, the olive oil market in Japan is heavily differentiated. There are various products of different quality and, hence, different prices, because of which consumers are now able to choose a price that reflects quality. Consumers may exhibit their purchase behavior not only in quantity but also quality. We aimed to analyze consumers' quantity choice and quality choice.

To this end, we aggregated the data into four groups by using two variables: the median of the price and the median of the purchase volume. By categorizing the consumers into four groups using these variables, we have four types of consumers that are different from the viewpoint of preferred olive oil quality, which is shown by the price they paid for it and the volume they consumed.

The data were analyzed using two types of models, logistic regression and classification models. Regarding the logistic model, the multinomial logit model was chosen. As for the classification models, the decision tree and random forest models were chosen. The purpose of the classification analysis can be either to produce an accurate classifier or to uncover the

predictive structure of the problem (Breiman et al., 1984). Our aim is the latter. By using the results of each of these models, we calculated the prediction accuracy to compare their efficiency.

2. Data

Scanner data constitute a source of product-specific information (Nayga and Oral, 1994). The data we use for this study represent the purchase history of olive oil from 5,197 consumers in Japan. These data were provided by Intage Inc., a Japanese marketing company. The data cover all of Japan through a period of two years (from August 2010 to July 2012).

The data were divided into four groups of consumers, by using the medians of the average purchase price and average volume of olive oil,[2] as shown in Table 4-1.

Each of these groups represents one type of consumer. The four types are a combination of the following: LV (consumed volume lower than the median), LP (price lower than the median), HP (price higher than the median), HV (consumed volume higher than the median). As such the four groups are: LVLP, HVLP, LVHP and HVHP.

[2] In this chapter, we took into account the quality of olive oil (which is reflected through the price).

Table 4-1: Grouping method for consumption types

Types	LVLP (Group 1)	LVHP (Group 2)	HVLP (Group 3)	HVHP (Group 4)
average volume (ml)	< 400	< 400	> 400	> 400
average price (yen/100ml)	< 101.08	> 101.08	< 101.08	> 101.08
rate of consumers	16.89%	34.62%	33.12%	15.37%
median consumption (ml) in group	227.5	135	928	717
median price / 100ml in group	85.5	132.5	78.49	118.52

Source: Compiled by the author

Table 4-2: Descriptive statistics of the independent variables

Variable name	Interval	Mean	Variance
Age [(1)]	1 ~ 10	7.102	4.77
Household income [(2)]	1 ~ 5	2.776	2.2
Household size	1 ~ 6	2.881	1.585
Children	0 or 1	0.402	0.424
Job 1: executive	0 or 1	0.014	0.0144
Job 2: regular employee	0 or 1	0.061	0.058
Job 3: part-time employee	0 or 1	0.03	0.03
Job 4: unemployed	0 or 1	0.17	0.141
North	0 or 1	0.05	0.048
East	0 or 1	0.064	0.0604
West	0 or 1	0.3552	0.229
South	0 or 1	0.165	0.138
Married men	0 or 1	0.175	0.144
Single men	0 or 1	0.072	0.066
Married women	0 or 1	0.69	0.214
Single women	0 or 1	0.064	0.06

[(1)] Age variable increases by increments of 5 years. 1: under 19, 2: 20–24, …, 10: over 60 years old

[(2)] Household income: 1: under 3.99 million yen, 2: 4–5.49 million yen, …, 5: over 9 million yen

Source: Compiled by the author

3. Methodology

We adopted three analytical models to understand olive oil consumption type and identify consumers' characteristics affecting olive oil consumption: multinomial logit, decision tree, and random forest. These models were applied to a multitude of subsamples for comparing the average error rate of the models.

3.1. Multinomial logit model

According to Train (2003), the easiest and most used discrete choice model is the multinomial logit model (MNL). This may be due to the fact that the "formula for choice probabilities takes a closed form and is readily interpretable" (Train, 2003). This model is used to explain discrete choices and is based on McFadden's random utility theory (RUT). The RUT states that the "utility can be expressed as the sum of a systematic (deterministic) component V_{ij}, which is expressed as a function of the attributes presented (consumers' demographic characteristics in our case), and a random (stochastic) component ε_{ij}". McFadden, (1973) has stated that "a study of choice behavior is described by the following three factors: (1) the objects of choice and sets of alternatives available to decision-makers, (2) the observed attributes of decision-makers, and (3) a model of individual choice and behavior and distribution of behavior patterns in the population". However, for this model, the choices (the dependent variable) are the types of consumers, that is, the category to which consumer i belongs. As explained above, the groups were created based on the volume consumed and the price of olive oil. The demographic characteristics are used as independent variables.

$$U(LVLP) = \alpha_1 + \beta_1 x_{ij} + \varepsilon_1 \quad (1)$$

$$U(LVHP) = \alpha_2 + \beta_2 x_{ij} + \varepsilon_2 \quad (2)$$

$$U(HVLP) = \alpha_3 + \beta_3 x_{ij} + \varepsilon_3 \quad (3)$$

$$U(HVHP) = 0 \quad (4)$$

(1), (2), (3), and (4) represent the model used in the analysis. Each of the equations in the above model represents a group. α represents the intercept, β_1 represents the coefficients of group 1 (LV/LP) consumers and x_{ij} represent the characteristics j of consumer i. As such, we have an equation for each group.

- Marginal effects:

$$\frac{\partial P_{in}}{\partial x_i} = \frac{\partial(e^{V_{in}}/\sum_h e^{V_{ih}})}{\partial x_i}$$

$$= P_{in}\left(\frac{\partial V_{in}}{\partial x_i} - P_{in} \times \frac{\partial V_{in}}{\partial x_i}\right) - P_{in}P_{ic}\frac{\partial V_{ic}}{\partial x_i} - P_{in}P_{ik}\frac{\partial V_{ik}}{\partial x_i}$$

$$= P_{in}(\beta_{xn} - P_{in}\beta_{xn}) - P_{in}P_{ic}\beta_{xc} - P_{in}P_{ik}\beta_{xk} - P_{in}P_{il}\beta_{xl}$$

Where

$$P_{in} = \frac{e^{V_{in}}}{\sum_{h\in\{1,2,3,4\}} e^{V_{ih}}}$$

3.2. Decision tree model

Du and Zhijun, (2002) explained that "most decision tree classifiers (e.g., CART and C4.5) perform classification in two phases: tree building and tree pruning. In the former, the decision tree model is built by repeatedly splitting the data set based on an optimal criterion until all or most of the records belonging to each of the partitions bear the same class label. To improve generalization of a decision tree, the latter is used to prune the leaves and branches responsible for classification of a single or very few data vectors ". Decision trees have certain advantages in that they are extremely easy to visualize and interpret as well as particularly fast in classifying data. The decision tree allows us to model the explanatory variable Y in function of the observed values of descriptive variables (X) within a data set D = (X, Y), where $x = (x1, ..., x_j)$ is a set of j descriptive variables (for instance, sociodemographic characteristics) associated with the individual (Adriana et al., 2013). In our case, the Y value is the group number (from group1 to group4) the consumer i belongs to and the x_{ij} value represents the socio-demographic characteristics j of the consumers (age, gender, region, etc.).

The decision tree model uses the Gini impurity measure (Gini index) to construct the trees. The Gini index shows the importance of the variables used to create the tree model. Du and Zhijun (2002) have defined the formula to calculate the Gini index for a data set S as follows:

$$Gini\ (S) = 1 - \sum_{j=1}^{m} P_h^2$$

Where P_j is the relative frequency of group h in S. Based on the Gini index of S, we can calculate the information gain for each attribute. For instance, we will partition the data by using an attribute x; then, the gain would be calculated as follows:

$$Gain\ (S, x) = Gini\ (S) - \sum_{v \in x_j} (\frac{|S_v|}{|S|} * Gini(S_v))$$

Where v represents any possible values of attribute x_j; S_v is the subset of S for which attribute x_j has value v; | S_v | is the number of elements in S_v; | S | is the number of elements in S.

3.3. Random forest model

Random forest is an ensemble learning method for classification and regression. While in the previous model, only one tree is created, random forest algorithms create an ensemble of decision trees using randomization. The random forest uses the same technique in creating each of those trees, that is, the Gini index. Each individual tree in the random forest gives a class prediction, and the class with the most votes becomes the model's prediction.

Decision trees will often overfit the data, unless some regularization methods, such as pruning or imposing a minimum number of training samples per leaf, are used (Vieira, 2016). The random forest technique avoids the overfitting problem when using the decision tree model. Even if the trees in the forest are grown without pruning, the fact that the classifiers' output depends on the entire set of trees and not on a single tree, the risk of overfitting is considerably reduced (Vieira, 2016).

Liaw and Wiener, (2002) have defined the algorithm of the random forest as follow:

i. "Draw n_{tree} bootstrap samples from the original data. n_{tree} represents the number of subsamples that the random forest creates and ultimately the number of trees.
ii. For each of the samples, the model grows an unpruned classification or a regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, it randomly samples m_{try} of the predictors (variables that model will use to create the tree) and chooses the best split from among those

variables. (Bagging can be thought of as the special case of random forests obtained when m_{try} = p, the number of predictors.)

iii. Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification, average for regression).

An estimate of the error rate can be obtained, based on the training data, as follows:

i. At each bootstrap iteration, predict the data not in the bootstrap sample ("out-of-bag" or OOB data) using the tree grown with the bootstrap sample.

ii. Aggregate the OOB predictions. (On the average, each data point would be out-of-bag around 36% of the time, so aggregate these predictions.) Calculate the error rate and call it the OOB estimate of error rate."

3.4. Sampling method to compare performances of the three models

We used the sampling procedure implemented by Jin (20019). Figure 4-1 explains this procedure to evaluate prediction performances in the three models. The testing set (20%) was chosen from the original sample randomly. The remainder of the sample was considered as the training set. The model parameters were estimated with the training set. Thereafter, the prediction for testing data was made with the estimated model.

The prediction of group for individuals was conducted as follows:

i. Using the results of the estimation and a function in R called "prediction," we predict the group number consumer i belongs to.

ii. We compare the results of this prediction with the actual group number of consumer i in the testing data.

iii. Finally, we calculate the percentage of correct predictions to determine the accuracy of each model.

An error rate was calculated by deducting the prediction accuracy, which was obtained by comparing the predicted value with the actual value, from one. This procedure was replicated 5,000 times. Therefore, 5,000 simulated error rates were obtained.

4. Results and discussion

4.1. Empirical results

The aim of this research is twofold: first, to estimate individual characteristics to understand how they affect the consumption types of consumers, and second, to compare the efficiency of three models (multinomial logit, decision tree, and random forest). Table 4-3 summarizes the results of the first part of our study. The values in the logit model's column represent the socio-demographic variables' coefficient. However, the values in the decision tree and random forest model columns represent the importance of these variables in the classification procedure.

According to the random forest model, age, household income, and family size are the variables that mostly affected the classification.

Unlike the random forest (RF) model, in the decision tree (DT) model, family size was the most important, followed by household income and age.

The results of the logit model are slightly different from those of the other two models. Indeed, the results of the logit model help us to understand how each of the socio-demographic characteristics affects olive oil consumption.

The age variable affects all types of consumers. In fact, according to the results of the logit model, LVLP and LVHP consumers tend to be young, while HVLP and HVHP categories are primarily older consumers. This result matches our hypothesis of how age affects the volume consumed and price of olive oil.

Table 4-4 lists the marginal effects of the independent variables on the grouping. Only "Region west" and "Region south" have affected the probability of group 1 (LVLP). Both these variables had a positive effect. LVHP was affected by many variables. While family size and job (regular) increased the probability of a consumer belonging to this group, south, and the "Marriage X Gender" variables all have a negative effect on that probability. Regarding HVLP, both age and south variables have a positive effect on its probability. Finally, HVHP's probability was negatively affected by family size and west and south.

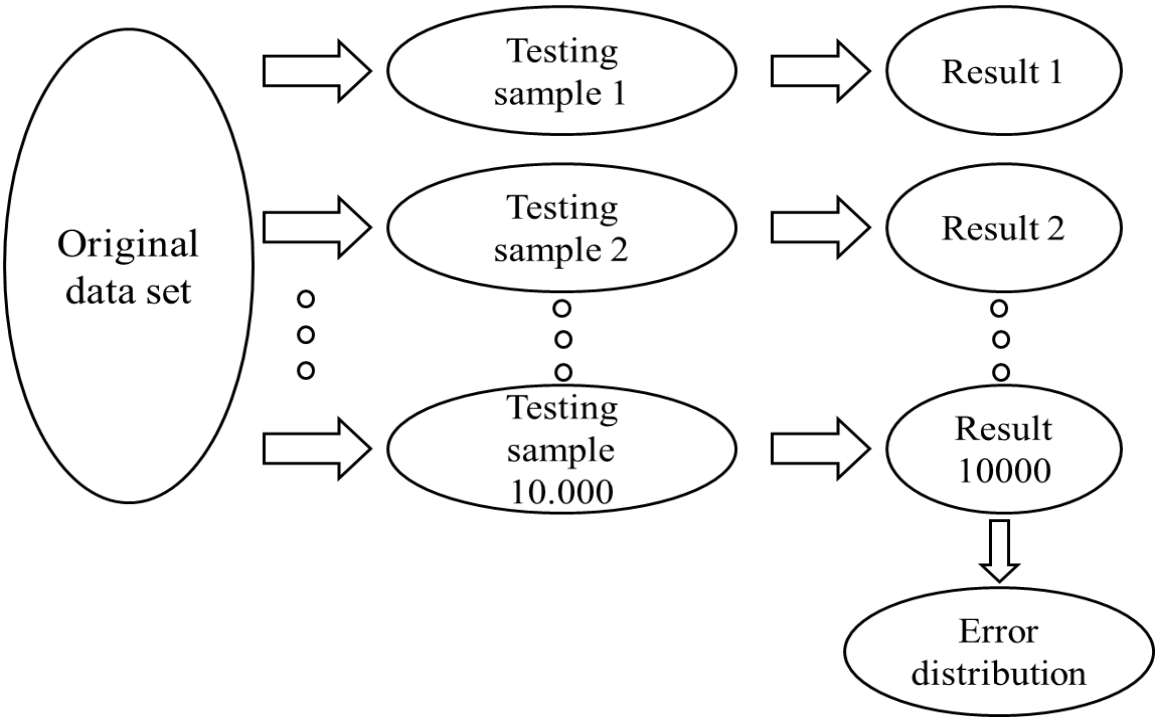


Figure 4-1: Sampling simulation procedure

Source: Compiled by the author

Table 4-3: Estimation results

explanatory variable		Logit model (base: HVHP)			Decision tree model	Random forest model
		LVLP	LVHP	HVLP		
Age		-0.06*	-0.09***	0.01	78.04	217.5
Children		0.21*	0.06	0.09	40.7	68.71
Family size		0.11 #	0.09#	0.08#	130.8	156.22
Household income		-0.031	-0.036	-0.015	89	179.57
Job	job1 (Executive)	0.16	0.11	0.23	73.1	122
	Job 2 (regular)	0.16	0.42**	0.16		
	Job 3(part time)	0.033	0.134	0.032		
Region	region N	0.16	0.023	0.12	39.42	118.7
	region W	0.73***	0.26*	0.24*		
	region S	0.64***	0.055	0.5**		
Marriage X gender	Married man	0.092	-0.46*	0.16	32.55	68.9
	Single man	0.04	-0.12	0.28		
	Married woman	-0.28	-0.7**	0.11		

Note1) ***, **, *, #: significance at 0.001, 0.01, 0.05 and 0.1 respectively

Note 2) The values in the logit model column are coefficient; the others are the mean decrease in gini

Source: Compiled by the author

Table 4-4 : Results of the Multinomial Logit Model regression (Marginal effects and their conf. int.)

explanatory variable		Logit model (base: HVHP) (Estimated coefficients)											
		LVLP	Conf. int.		LVHP	Conf. int.		HVLP	Conf. int.		HVHP	Conf. int.	
Intercept		0.0098	-0.073	0.094	0.2365	0.1705	0.3074	-0.162	-0.297	-0.033	-0.085	-0.234	0.0723
Age		-0.005	-0.012	0.002	-0.015	-0.021	-0.008	0.016	0.0061	0.0264	0.0037	-0.009	0.0162
Children		0.0281	-0.002	0.054	-0.011	-0.037	0.012	-2E-04	-0.042	0.0415	-0.016	-0.066	0.0359
Family size		0.0075	-0.008	0.025	0.0201	0.0061	0.0333	0.0082	-0.017	0.034	-0.035	-0.067	-0.005
Household income		-0.007	-0.017	0.002	-0.006	-0.015	0.003	0.008	-0.006	0.0221	0.0047	-0.012	0.0209
Job	job1 (Executive)	0.0032	-0.059	0.061	-0.002	-0.054	0.049	0.0463	-0.035	0.1248	-0.046	-0.144	0.044
	Job 2 (regular)	-0.012	-0.055	0.03	0.0603	0.0267	0.0985	-0.004	-0.065	0.0577	-0.044	-0.12	0.0268
	Job 3(part time)	-0.011	-0.045	0.024	0.018	-0.01	0.0481	0.0037	-0.044	0.0543	-0.011	-0.067	0.0434
Region	region N	0.0113	-0.037	0.06	-0.011	-0.047	0.0279	0.0098	-0.054	0.072	-0.009	-0.083	0.0606
	region W	0.0836	0.048	0.116	0.0045	-0.026	0.0341	0.0027	-0.044	0.0472	-0.09	-0.149	-0.03
	region S	0.0943	0.045	0.142	-0.053	-0.101	-0.003	0.1036	0.029	0.1761	-0.145	-0.243	-0.044
Marriage X gender	Married man	-0.055	-0.128	0.018	-0.106	-0.172	-0.041	0.0969	-0.03	0.2306	0.0602	-0.082	0.2093
	Single man	-0.067	-0.147	0.018	-0.079	-0.142	-0.011	0.1176	-0.013	0.2467	0.0273	-0.118	0.1747
	Married woman	-0.069	-0.132	9E-04	-0.137	-0.192	-0.079	0.1161	-0.009	0.2373	0.0838	-0.044	0.2252

Source: Compiled by the author

The children variable shows that LVLP consumers tend to have more children than HVHP consumers. The results of the region variable show that HVHP consumers are mainly present in the eastern region of Japan. All results of the three models indicated that age and family size affected consumption type. In some respects, each of these models demonstrated different results. For example, for the RF model, age was the most important factor affecting the consumption behavior toward olive oil in Japan, followed by household income and family size. Meanwhile, in the DT model, family size was the most important demographic characteristic, followed by household income and age. Finally, in the multinomial logit model, age, children, job, region, and the cross variables of marriage and gender were the only significant variables. Region and family size were the only variables that affected all types of consumers.

4.2. Methodological results

The second part of our results demonstrates the comparison of the three models by comparing the error of each of these models.

Figure 4-2 represents the error distribution of the multinomial logit (MNL), DT, and RF models. This figure shows that the DT model has the lowest average error, while the multinomial logit model had the highest one. The error rate was computed as follows:

$$Error\ rate = 1 - accuracy\ rate$$

Based on this figure, the DT model is the most efficient model in classifying the consumers, though the RF model is not far behind. In fact, Table 4-4 shows that no significant difference exists between the error terms of the DT model and the RF model. As per Table 4-4, the only significant error is between the decision tree and the multinomial logit models. For the error terms in the four types of consumers, there was only a significant difference between the RF and the multinomial logit models for type 1 (LVLP).

We note that the average error for type 1 and type 2 consumers is quite high; in fact, for type 4 consumers, the average error was 1. These high values of the average error can be

Table 4-5: Error term by model and consumer type

	Logit model	Decision tree model	Random forest model
Total	66.2%[0.63; 0.69]*	63.04% [0.6;0.66] *	63.4% [0.6; 0.66]
Type 1 (LVLP)	99.92%[0.99; 1]**	97.04%[0.92; 1]	96.3% [0.93; 0.99]**
Type 2 (LVHP)	53.04% [0.47; 0.59]	46.89% [0.36; 0.58]	47.31%[0.4; 0.54]
Type 3 (HVLP)	47.01%[0.4; 0.53]	45.69%[0.33; 0.57]	47.02% [0.4; 0.54]
Type 4 (HVHP)	100% [1; 1]	99.25%[0.96; 1]	98.6% [0.96; 1]

Note 1) The 95% confidence interval was shown in the bracket.

Note 2) * there is a significant difference at 10% significance level.

Note 3) ** there is a significant difference at 2% significance level.

Source: Compiled by the author

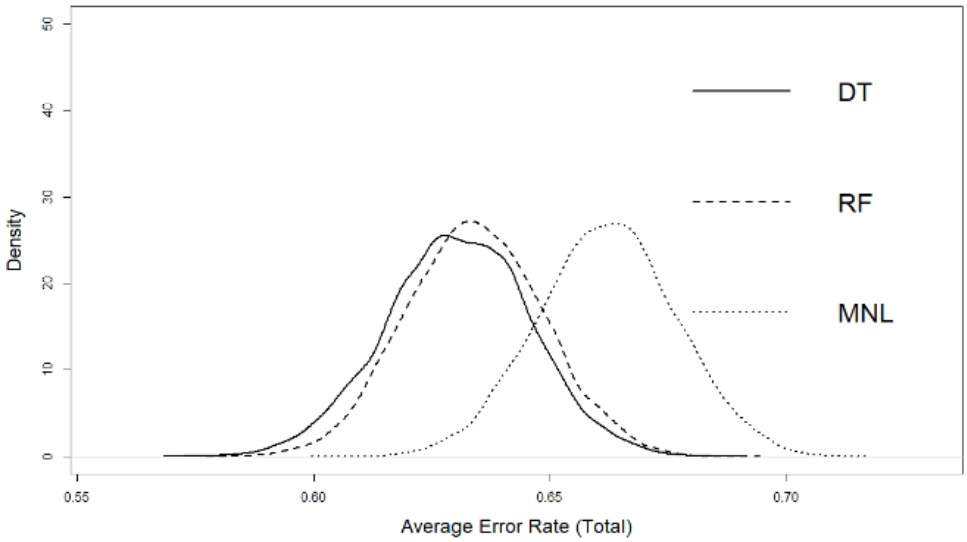


Figure 4-2: Error term distribution

Source: Compiled by the author

explained by the fact that the number of consumers that belong to type 1 and type 4 is quite small compared to the other types (2 and 3).

5. Conclusion

We used three models to study the characteristics of olive oil consumption and compared the prediction accuracy of the models.

Regarding type 1 (LVLP) consumers, age, job, and region seem to determine this group. Type 1 consumers are young with children and are mostly situated in western and southern Japan. Type 2 (LVHP) consumers are young with a bigger family size than HVHP and live in western Japan. Additionally, they have regular jobs and are mostly unmarried.

Finally, regarding type 3 (HVLP) consumers, only family and region seem to affect them. This group of consumers are mostly living in southern and western Japan and have a big family size. Older consumers tend to purchase more than the younger consumers. A small family might have type 4 (HVHP) consumers. Household income may affect consumer behavior. The higher the income of a household, the higher is the probability that the household would have HVHP consumption. However, the relation between household income and consumption type is not linear.

For the second part of our study, that is, the prediction accuracy of the models, a significant difference existed between the decision tree and the multinomial logit models. In terms of accuracy, the DT model had the highest accuracy, followed by the RF model, while the multinomial logit model had the lowest accuracy. Regarding the average error in each group of consumers, type 1 and type 4's errors were extremely high, even reaching 1 in type 4. This may be because these two types are not common, since the number of consumers belonging to these two types is quite low compared to the other two types. However, it has yet to be proven, and additional research should be conducted to prove this point.

On the one hand, the decision tree is the most suitable model for classification, since it has the highest accuracy. On the other hand, the multinomial logit model is the most appropriate one to explain how each of the socio-demographic variables affects consumption.

There is still much to explore regarding Japanese olive oil consumers, and further research should be conducted in this domain. Concerning this study, only the socio-demographic characteristics of the consumers were used to explain the consumption behavior. By relying solely on these consumer characteristics, we were able to make predictions with an accuracy of almost 37%, which is 12% higher than if we classify consumers randomly without relying on anything. By including more variables in this model, we may be able to increase the accuracy even more.

While the decision tree was considered the most suitable model in classifying the consumers in this case, this finding cannot be generalized; in fact, according to Breiman (2001), the RF model is like an upgrade of the decision tree model.

Other models, such as the AIDS model, can also be used to analyze scanner data. We will consider the AIDS model and compare it with these models, or others, in future research.

Chapter 5：Effects of the Health Consciousness on the Olive Oil Consumption in Japan

This work has not been published yet so I can not make it public.

Chapter 6 : Conclusion

In this thesis, we first studied the oil market in Japan. We used scanner panel data along with attitudinal data gathered through a survey. Both datasets were used simultaneously to first assess the health consciousness effect on oil consumption and then on olive oil consumption in Japan.

Though oil is not a part of the traditional Japanese diet, its consumption has considerably risen during the last few decades. In fact, there has been an increase in western eating habits in Japan. In the past few years, the Japanese market has seen a new trend of healthy oils, also known as superfood, such as coconut oil or flaxseed oil. Aside from these healthy oils, olive oil, though not a superfood but nonetheless healthy, has gained an incredible popularity among Japanese consumers. Indeed, the retail sales value of olive oil is only second to Canola oil in Japan. Since production of olive oil in Japan is low, the country mainly relies on imports.

Additionally, we divided health consciousness into 3 categories: food, sports, and lifestyle.

Food: This category expresses consumers' attitudes towards food consumption. For instance, consumers would try to limit their sugar or salt intake, or they would avoid eating too much meat, etc. While this category directly relates to our topic of study (i.e., oil consumption,) it is not the only important factor in determining the health conscious of consumers, nor is it in direct correlation with oil consumption.

Sports: This category shows consumers' attitude towards physical activities. The more they practice sports on a regular basis, the higher their sense of health consciousness should be. Though a sport isn't always practiced with the intention of staying healthy, it still remains one of the main reasons why one practices it. Whether they practice sports to lose weight or gain muscles, it is very beneficial for one's health.

Lifestyle: This category depicts how consumers try to have a healthy lifestyle by getting enough sleep or having regular health check-ups at the hospital. Additionally, they try to avoid stress to keep a healthy mind.

Aside from health consciousness, the models also incorporated variables pertaining to the health status of consumers. To this end, diseases such as diabetes, cholesterol, high blood pressure, and being overweight, which have a relation with the health benefits of olive oil, were considered.

In our first study on health consciousness' effect on oil purchase choice among Japanese consumers (chapter 3), we used the NBD-Dirichlet model which is composed of NBD and DMD. NBD was used to analyze the oil consumption as a whole and DMD was used to analyze the choice of type of oil to buy. Among the demographic characteristics, only age and household size were positively significant in the negative binomial model results. Regarding the variables pertaining to health consciousness, they had different effects on oil consumption. While sports negatively affected consumption, lifestyle and food had positive effect on oil consumption. Last but not the least, only cholesterol had a significant impact on choice to buy oil; indeed, consumers suffering from high levels of cholesterol have lower consumption of oil.

As for the Dirichlet model results, it appears that all 3 variables pertaining to health consciousness have positively affected consumption of healthy oils (Olive oil, Coconuts oil, Linseed oil, Perilla oil) while negatively affecting Canola oil consumption which is considered unhealthy. For the health status variables, consumers with diabetes seem to reduce their consumption of all the types of oil, healthy and unhealthy, while consumers suffering from high blood pressure and high cholesterol levels increase their consumption of healthy oils and reduce that of unhealthy oils. It is important to note that in this study, only the frequency of purchase was taken into account. The volume and price of oils wasn't considered since the NBD-Dirichlet model can only be used with count data. Although the volume of consumed oil increases with the number of bottles of oil bought, the increase is less than pro rata.

In our second study (chapter 4), we analyzed olive oil consumption patterns by using 3 different models: Multinomial Logit, Decision Tree, and Random Forest. First, the consumers were split in 4 categories depending on the quantity and price of the olive oil they consume. The results showed that age, family size, and region were important factors in describing the type of olive oil consumers. For instance, age negatively affected type 2 consumers. Family size however had a positive impact on type 1, 2, and 3 consumers. Regarding the prediction accuracy of the 3 models, decision tree and random forest had the highest accuracy while multinomial logit had the lowest accuracy.

In the final study (Chapter 5), we studied the health consciousness effect on olive oil consumption by using a Hierarchical Bayesian model. The variables used in this chapter were the same as those in the first study. The estimation results corroborate the results of the first

study which showed that health conscious consumers would buy more healthy oils. Indeed, both sports and food positively affected olive oil consumption, signifying that consumers who follow a healthy diet and regularly practice sports have a higher consumption of olive oil. The health status variables showed that only diabetes had a significant effect on olive oil consumption with those suffering from diabetes buying less olive oil than the others.

Health consciousness has played an important role in olive oil consumption in Japan. Unlike the Mediterranean region where olive oil is a traditional product, it is relatively new in Japan and even though it isn't a part of the traditional Japanese diet, it has gained tremendous popularity. We believe that this was because of the health aspects of this product. In fact, based on the results of our study, health conscious consumers tend to consume more olive oil than those who aren't.

References

1. Adhikari, P., Hwang, K. T., Park, J. N., & Kim, C. K. (2006). Policosanol content and composition in perilla seeds. Journal of agricultural and food chemistry, 54(15), 5359-5362.
2. Aki Vehtari, Prior Choice Recommendations, Github, (24/07/2019), https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.
3. Alexandre Perrault, "Sector Trend Analysis, Vegetable Oils in Japan", Ministry of Agriculture and Agri-food Canada, http://www.agr.gc.ca, 2017 (9/19/2019).
4. Allenby, G. M., Arora, N., & Ginter, J. L. (1998). On the heterogeneity of demand. *Journal of Marketing Research*, *35*(3), 384-389.
5. Allenby, G. M., Bradlow, E. T., George, E. I., Liechty, J., & McCulloch, R. E. (2014). Perspectives on Bayesian methods and big data. Customer Needs and Solutions, 1(3), 169-175.
6. Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. Journal of econometrics, 89(1-2), 57-78.
7. Alonso, A., de la Fuente, C., Martín-Arnau, A. M., de Irala, J., Martínez, J. A., & Martínez-González, M. Á. (2004). Fruit and vegetable consumption is inversely associated with blood pressure in a Mediterranean population with a high vegetable-fat intake: the Seguimiento Universidad de Navarra (SUN) Study. British Journal of Nutrition, 92(2), 311-319.
8. Aprile, M. C., Caputo, V., & Nayga Jr, R. M. (2012). Consumers' valuation of food quality labels: the case of the European geographic indication and organic farming labels. International Journal of Consumer Studies, 36(2), 158-165.
9. Banerjee, A. K., & Bhattacharyya, G. K. (1981). Use of Bayesian analysis of semi-Markov process models to study consumer buying behavior. American Journal of Mathematical and Management Sciences, 1(2), 109-137.
10. Bassi, F. (2011). The Dirichlet Model: Analysis of a Market and Comparison of Estimation Procedures. *Marketing Bulletin*, *22*.
11. Byun, H., & Lee, C. Y. (2017). Analyzing Korean consumers' latent preferences for electricity generation sources with a hierarchical Bayesian logit model in a discrete choice experiment. Energy Policy, 105, 294-302.
12. Caporale, G., Policastro, S., Carlucci, A., & Monteleone, E. (2006). Consumer expectations for sensory properties in virgin olive oils. *Food quality and preference*, 17(1-2), 116-125.
13. Carandang, E. V. (2008). Health benefits of virgin coconut oil. *INDIAN COCONUT JOURNAL-COCHIN-*, *38*(9), 8.
14. Carlin, B. P., & Louis, T. A. (2010). Bayes and empirical Bayes methods for data analysis. Chapman and Hall/CRC.
15. Casini, L., Rungie, C., & Corsi, A. M. (2009). How loyal are Italian consumers to wine attributes?. *Journal of Wine Research*, *20*(2), 125-142.
16. Cavallo, C., & Piqueras‐Fiszman, B. (2017). Visual elements of packaging shaping healthiness evaluations of consumers: The case of olive oil. Journal of sensory studies, 32(1), e12246.
17. Chandukala, S. R., Dotson, J. P., Brazell, J. D., & Allenby, G. M. (2011). Bayesian analysis of hierarchical effects. Marketing Science, 30(1), 123-133.
18. Chan-Halbrendt, C., Zhllima, E., Sisior, G., Imami, D., & Leonetti, L. (2010). Consumer preferences for olive oil in Tirana, Albania. *International Food and Agribusiness Management Review*, *13*(1030-2016-82871).

19. Chen, M. F. (2009). Attitude toward organic foods among Taiwanese as related to health consciousness, environmental attitudes, and the mediating effects of a healthy lifestyle. *British food journal*, *111*(2), 165-178.
20. Chern, Wen S., Edna T. Loehman, and Steven T. Yen. (1995). Information, health risk beliefs, and the demand for fats and oils. *The Review of Economics and Statistics*: 555-564.
21. Covas, M. I., Konstantinidou, V., & Fitó, M. (2009). Olive oil and cardiovascular health. *Journal of cardiovascular pharmacology*, *54*(6), 477-482.
22. Covas, M. I., Ruiz-Gutiérrez, V., De La Torre, R., Kafatos, A., Lamuela-Raventós, R. M., Osada, J., ... & Visioli, F. (2006). Minor components of olive oil: evidence to date of health benefits in humans. *Nutrition Reviews*, *64*(suppl_4), S20-S30.
23. DebMandal, M., & Mandal, S. (2011). Coconut (Cocos nucifera L.: Arecaceae): in health promotion and disease prevention. *Asian Pacific Journal of Tropical Medicine*, *4*(3), 241-247.
24. Delgado, C., & Guinard, J. X. (2010). How do consumer hedonic ratings for extra virgin olive oil relate to quality ratings by experts and descriptive analysis ratings?, *Food Quality and Preference*, 22(2), 213-225.
25. Dekhili, S., Sirieix, L., & Cohen, E. (2011). How consumers choose olive oil: The importance of origin cues. *Food quality and preference*, *22*(8), 757-762.
26. Ehrenberg, A. S. (1959). The pattern of consumer purchases. Journal of the Royal Statistical Society: Series C (Applied Statistics), 8(1), 26-41.
27. Ellison, B., Lusk, J. L., & Davis, D. (2013). Looking at the label and beyond: the effects of calorie labels, health consciousness, and demographics on caloric intake in restaurants. International Journal of Behavioral Nutrition and Physical Activity, 10(1), 21.
28. Einav, L., Leibtag, E., & Nevo, A. (2010). Recording discrepancies in Nielsen Homescan data: Are they present and do they matter?. QME, 8(2), 207-239.
29. Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. Journal of consulting and clinical psychology, 43(4), 522.
30. Gámbaro, A., Ellis, A. C., & Prieto, V. (2013). Influence of subjective knowledge, objective knowledge and health consciousness on olive oil consumption—A case study.
31. Gebauer, S. K., Psota, T. L., Harris, W. S., & Kris-Etherton, P. M. (2006). n− 3 fatty acid dietary recommendations and food sources to achieve essentiality and cardiovascular benefits. *The American journal of clinical nutrition*, *83*(6), 1526S-1535S.
32. Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian analysis, 1(3), 515-534.
33. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*, 3rd edition. Chapman and Hall/CRC.
34. Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis jumping rules. Bayesian statistics, 5(599-608), 42.
35. Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 555.
36. Goodhardt, G. J., Ehrenberg, A. S., & Chatfield, C. (1984). The Dirichlet: a comprehensive model of buying behaviour. Journal of the Royal Statistical Society. Series A (General), 621-655.
37. Gorzynik-Debicka, M., Przychodzen, P., Cappello, F., Kuban-Jankowska, A., Marino Gammazza, A., Knap, N., ... & Gorska-Ponikowska, M. (2018). Potential health benefits of olive oil and plant polyphenols. *International journal of molecular sciences*, *19*(3), 686.

38. Gould, S. J. (1988). Consumer attitudes toward health and health care: A differential perspective. Journal of Consumer Affairs, 22(1), 96-118.
39. Gunstone, F. (Ed.). (2011). *Vegetable oils in food technology: composition, properties and uses*. John Wiley & Sons.
40. Hasler, C. M. (1998). Functional foods: their role in disease prevention and health promotion. *FOOD TECHNOLOGY-CHAMPAIGN THEN CHICAGO-*, *52*, 63-147.
41. Henningsen, A. (2010). Estimating censored regression models in R using the censReg Package. R package vignettes collection, 5(2), 12.
42. Hong, H. (2009). Scale development for measuring health consciousness: Re-conceptualization. that Matters to the Practice, 212.
43. Ibrahim, J. G., & Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. Journal of the American Statistical Association, 86(416), 981-986.
44. Inarejos-García, A. M., Santacatterina, M., Salvador, M. D., Fregapane, G., & Gómez-Alonso, S. (2010). PDO virgin olive oil quality—Minor components and organoleptic evaluation. *Food Research International*, 43(8), 2138-2146.
45. Jarvis, W., Rungie, C., & Lockshin, L. (2003). Analysing wine behavioural loyalty (Doctoral dissertation, University of South Australia).
46. Jayanti, R. K., & Burns, A. C. (1998). The antecedents of preventive health care behavior: An empirical study. Journal of the academy of marketing science, 26(1), 6-15.
47. Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 186(1007), 453-461.
48. Jolliffe, I. T. (1986). Principal components in regression analysis. In Principal component analysis (pp. 129-155). Springer, New York, NY.
49. Jussaume Jr, R. A., & Judson, D. H. (1992). Public Perceptions about Food Safety in the United States and Japan 1. *Rural Sociology*, *57*(2), 235-249.
50. Kass, R. E., & Wasserman, L. (1996). Formal rules for selecting prior distributions: A review and annotated bibliography. Journal of the American Statistical Association, 435, 1343-1370.
51. Kasteridis, P., & Yen, S. T. (2012). US demand for organic and conventional vegetables: a Bayesian censored system approach. Australian Journal of Agricultural and Resource Economics, 56(3), 405-425.
52. Kearney, J. (2010). Food consumption trends and drivers. Philosophical transactions of the royal society B: biological sciences, 365(1554), 2793-2807.
53. KIM, S. R., & Chern, W. S. (1999). Alternative measures of health information and demand for fats and oils in Japan. Journal of consumer affairs, 33(1), 92-109.
54. Kraft, F. B., & Goodell, P. W. (1993). Identifying the health-conscious consumer. *Marketing Health Services*, *13*(3), 18.
55. Krystallis, A., & Ness, M. (2005). Consumer preferences for quality foods from a South European perspective: A conjoint analysis implementation on Greek olive oil. *International Food and Agribusiness Management Review*, *8*(1030-2016-82535), 62-91.
56. Kurowska, E. M., Dresser, G. K., Deutsch, L., Vachon, D., & Khalil, W. (2003). Bioavailability of omega-3 essential fatty acids from perilla seed oil. *Prostaglandins, leukotrienes and essential fatty acids*, *68*(3), 207-212.

57. Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. R news, 2(3), 18-22.

58. López-Miranda, J., Pérez-Jiménez, F., Ros, E., De Caterina, R., Badimón, L., Covas, M. I., ... & de la Lastra, C. A. (2010). Olive oil and health: summary of the II international conference on olive oil and health consensus report, Jaén and Córdoba (Spain) 2008. Nutrition, metabolism and cardiovascular diseases, 20(4), 284-294.

59. Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). Computers & Geosciences, 19(3), 303-342.

60. Mai, R., & Hoffmann, S. (2012). Taste lovers versus nutrition fact seekers: how health consciousness and self‐efficacy determine the way consumers choose food products. Journal of Consumer Behaviour, 11(4), 316-328.

61. Maritz, J. S., & Lwin, T. (2018). Empirical bayes methods. Routledge.

62. Martínez-González, M. Á., & Sánchez-Villegas, A. (2004). The emerging role of Mediterranean diets in cardiovascular epidemiology: monounsaturated fats, olive oil, red wine or the whole pattern?. *European journal of epidemiology*, *19*(1), 9-13.

63. McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

64. Michaelidou, N., & Hassan, L. M. (2008). The role of health consciousness, food safety concern and ethical identity on attitudes and intentions towards organic food. International journal of consumer studies, 32(2), 163-170.

65. Mtimet N., Kiyokazu Ujiie, Kenichi Kashiwagi, Lokman Zaibet and Masakazu Nagaki (2011). The effects of Information and Country of Origin on Japanese Olive Oil Consumer Selection, EAAE 2011 Congress Change and Uncertainty Challenges for Agriculture, Food and Natural Resources.

66. Muth, M. K., Siegel, P. H., & Zhen, C. (2007). Homescan data description. ERA Data Quality Study Design.

67. Nagata, J., & Yamada, K. (2008). Foods with health claims in Japan. Food science and technology research, 14(6), 519-519.

68. Nagendra Prasad, M. N., Sanjay, K. R., Shravya Khatokar, M., Vismaya, M. N., & Nanjunda Swamy, S. (2011). Health benefits of rice bran-a review. *J Nutr Food Sci*, *1*(3), 1-7.

69. Natarajan, R., & Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. Journal of the American Statistical Association, 95(449), 227-237.

70. Natarajan, R., & McCulloch, C. E. (1998). Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference?. Journal of Computational and Graphical Statistics, 7(3), 267-277.

71. Nayga Jr, R. M., & Capps Jr, O. (1994). Tests of weak separability in disaggregated meat products. American Journal of Agricultural Economics, *76*(4), 800-808.

72. Newsom, J. T., McFarland, B. H., Kaplan, M. S., Huguet, N., & Zani, B. (2005). The health consciousness myth: implications of the near independence of major health behaviors in the North American population. Social Science & Medicine, 60(2), 433-437.

73. Owen, R. W., Giacosa, A., Hull, W. E., Haubner, R., Würtele, G., Spiegelhalder, B., & Bartsch, H. (2000). Olive-oil consumption and health: the possible role of antioxidants. The lancet oncology, 1(2), 107-112.

74. Pehowich, D. J., Gomes, A. V., & Barnes, J. A. (2000). Fatty acid composition and possible health effects of coconut constituents. *West Indian Medical Journal*, *49*(2), 128-133.

75. Popa, V. M., Gruia, A., Raba, D. N., Dumbrava, D., Moldovan, C., Bordean, D., & Mateescu, C. (2012). Fatty acids composition and oil characteristics of linseed (Linum Usitatissimum L.) from Romania. *Journal of Agroalimentary Processes and Technologies*, *18*(2), 136-140.

76. Roselli, L., Clodoveo, M. L., Corbo, F., & De Gennaro, B. (2017). Are health claims a useful tool to segment the category of extra-virgin olive oil? Threats and opportunities for the Italian olive oil supply chain. *Trends in Food Science & Technology*, *68*, 176-181.

77. Rossi, P. E., & Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science*, *22*(3), 304-328.

78. Rungie, C. (2003). How to estimate the parameters of the Dirichlet model using likelihood theory in Excel. Marketing Bulletin, 14(3), 1-9.

79. Rungie, C., Uncles, M., & Laurent, G. (2013). Integrating consumer characteristics into the stochastic modelling of purchase loyalty. European Journal of Marketing, 47(10), 1667-1690.

80. Saija, A., & Uccella, N. (2000). Olive biophenols: functional effects on human wellbeing. *Trends in Food Science & Technology*, *11*(9-10), 357-363.

81. Santosa, M., Abdi, H., & Guinard, J. X. (2010). A modified sorting task to investigate consumer perceptions of extra virgin olive oils. *Food Quality and Preference*, 21(7), 881-892.

82. Santosa, M., Clow, E. J., Sturzenberger, N. D., & Guinard, J. X. (2013). Knowledge, beliefs, habits and attitudes of California consumers regarding extra virgin olive oil. *Food research international*, *54*(2), 2104-2111.

83. Servili, M., & Montedoro, G. (2002). Contribution of phenolic compounds to virgin olive oil quality. *European Journal of Lipid Science and Technology*, *104*(9-10), 602-613.

84. Stan Development Team. (2019), Stan user's guide, https://mc-stan.org/docs/2_21/stan-users-guide-2_21.pdf.

85. Sugano, Michihiro. "Characteristics of fats in Japanese diets and current recommendations." *Lipids* 31.1Part2 (1996): S283.

86. Tai, S. H., & Tam, J. L. (1997). A lifestyle analysis of female consumers in greater China. Psychology & Marketing, 14(3), 287-307.

87. Takeshita, H. (1999). "Econometric analysis of health information impact on food consumption", *Journal of Rural Economics (Japan)* 71-2, 61-70.

88. Tonon, R. V., Grosso, C. R., & Hubinger, M. D. (2011). Influence of emulsion composition and inlet air temperature on the microencapsulation of flaxseed oil by spray drying. *Food Research International*, *44*(1), 282-289.

89. Tripathi, V., Abidi, A. B., Markerb, S., & Bilal, S. (2013). Linseed and linseed oil: health benefits-a review. *Int J Pharm Biol Sci*, *3*(3), 434-442.

90. Tuck, K. L., & Hayball, P. J. (2002). Major phenolic compounds in olive oil: metabolism and health effects. *The Journal of nutritional biochemistry*, *13*(11), 636-644.

91. Vlontzos, G., & Duquenne, M. N. (2014). Assess the impact of subjective norms of consumers' behaviour in the Greek olive oil market. *Journal of Retailing and Consumer Services*, 21(2), 148-157.

92. Wang, H. M. D., Kalwani, M. U., & Akçura, T. (2007). A Bayesian multivariate Poisson regression model of cross-category store brand purchasing behavior. Journal of Retailing and Consumer Services, 14(6), 369-382.

93. Wrigley, Neil, and Rita Dunn, (1984). "Stochastic panel-data models of urban shopping behavior: 1. Purchasing at individual stores in a single city", Environment and Planning A 16.5, 629-650.

94. Wrigley, Nigel, and Roger Dunn. (1984). Stochastic panel-data models of urban shopping behavior: 2. Multistore purchasing patterns and the Dirichlet model, Environment and Planning A 16.6, pp.759-778.

95. Wrigley, Neil, and Rita Dunn. (1984). Stochastic panel-data models of urban shopping behavior: 3. The interaction of store choice and brand choice, Environment and Planning A 16.9, pp.1221-1236.

96. Wrigley, Neil, and Richard Dunn. (1985). Stochastic panel-data models of urban shopping behavior: 4. Incorporating independent variables into the NBD and Dirichlet models, Environment and Planning A 17.3, pp.319-331.

97. Yang, R., & Berger, J. O. (1996). A catalog of noninformative priors (pp. 97-42). Institute of Statistics and Decision Sciences, Duke University.

98. Yang, S., Chen, Y., & Allenby, G. M. (2003). Bayesian analysis of simultaneous demand and supply. Quantitative marketing and economics, 1(3), 251-275.

99. Yen, Steven T., and Wen S. Chern. (1992). "Flexible demand systems with serially correlated errors: fat and oil consumption in the United States." *American Journal of Agricultural Economics* 74.3, 689-697.