Predicting environmentally friendly consumption behaviour and analyzing the effect of ecolabels on willingness-to-pay for organic food products

MASTER DATA SCIENCE & MARKETING ANALYTICS

Author:
Larissa Elena van Boxem

Student number:
477257

Supervisor: Dr. P. Wan

Co-reader: H. Deng

November 26, 2021

Abstract

In this paper, the effect of ecolabels on purchase behaviour and willingness-to-pay for organic food products is studied. Using survey data and existing consumer data from the Liss Panel, willingness-to-pay for 5 different types of organic food products is measured with the help of mixed logit models. In addition, consumers are classified as organic and non-organic consumers with random forests and support vector machines. The study finds that personal values and consumer demographics are very important predictors of organic consumption behaviour and that the majority of consumers is not prone to price changes in organic food products.

Contents

1 Introduction

It is hard to go around nowadays: the large number of organic food products, meat substitutes and healthy alternatives of our ordinary day-to-day products. Organic food products are assumed to be better for the environment because of their origin from organic farming practices, which involve the recycling of resources, the promotion of ecological balance, and the conservation of biodiversity. Ecolabeling has become an increasingly important tool to promote sustainable products around the world (Roheim et al., 2011). However, ecolabeled food products often come with a substantially higher price tag in comparison to their non-organic alternatives. Roheim et al. (2011) find a price premium of 14.2% on ecolabeled seafood products in the United Kingdom due to the higher costs related to the production and farming process of biological products.

To make sustainable farming more attractive to farmers, the Dutch government subsidizes farmers that apply more environmentally friendly farming practices (Rijksoverheid, 2021b). This subsidy is called the "green payment" and applies to farmers who, for example, grow different crops which promotes biodiversity, or farmers who devote 5% of their farmland to ecological practices. This payment is equal to €111.96 per hectare (Rijksdienst voor Ondernemend Nederland, 2020). The costs for switching from conventional to organic farming, however, are approximately €5,000 per hectare and the payback period is between 10 to 15 years (Mons, 2020). In 2020, 30% of the budgeted direct payments were assigned to this subsidy, totaling to approximately 220 million euros (Ministerie van Landbouw, Natuur en Voedselkwaliteit, 2019).

Despite the efforts of the government, subsidies are unable to cover enough of the higher costs, resulting in significant price premiums on organic products and the overall sector remaining small, especially in comparison to other countries. In 2018, only 3.2% of all farmland was assigned to organic farming practices (Rijksoverheid, 2021a). This percentage was 24.1% in Austria, 20.6% in Estonia, 20.3% in Sweden and approximately 15% in Italy. In 2019, the organic agricultural area grew by 7% relative to the year before (Rijksoverheid, 2021a). In 2020, the sector only grew by 2.5% (Skal, 2021).

This research studies the effect of price changes of organic and environmentally friendly food products on consumer preferences, and determines whether organic purchase behaviour can be predicted. It attempts to reveal consumer preferences that could possibly make organic farming practices more attractive. Furthermore, being able to predict consumer purchase behaviour with respect to organic food products could have important implications for marketing practices in stimulating environmentally friendly consumption behaviour. Given the two previously mentioned objectives, this research is built around two research questions:

1. *Is it possible to predict organic consumption behaviour, and what are the main determinants of organic consumption behaviour?*

2. *What is the effect of organic food labels on willingness-to-pay (WTP) for different types of food products?*

This research uses primary data collected from a survey distributed among 360 people in the Netherlands. In the first part of the analysis, the data are used to predict whether consumer data allows us to classify consumers into organic and non-organic consumers. This is done with the help of random forests classification models and support vector machines, after which the importance of the different variables on the classification is discussed. To get a general idea of possible important predictors of organic consumption behaviour and the effect size thereof, an ordered logistic regression is performed beforehand.

In the second part of the analysis, relating to the WTP for organic food products, a second data set is used to support the findings from the primary data. These data are collected from the Liss Panel Data Archive and aim to identify consumer heterogeneity with respect to morality in consumption decisions and perceptions of animal welfare. A mixed logit is used to analyze the effect of different food product attributes on purchase decisions and to identify WTP for organic food products.

The findings of this study could reveal new consumer preferences that would make organic farming practices more attractive and less risky for farmers who are considering switching to sustainable farming. In addition, being able to accurately predict organic consumption behaviour is of great value in order to target organic consumers effectively and avoid expensive and inefficient marketing practices. Furthermore, the results of this research could reveal new information on the price elasticity of organic products. If a small change in price points for organic food products results in a strong change in consumer preferences, slightly decreasing the price premium on organic products may be able to boost demand in such a way that farmers can compensate for the high costs of sustainable farming.

This study also carries scientific relevance as existing studies show that there is no one-size-fits-all policy: WTP for organic products differs per country (Galarraga Gallastegui, 2002) and research conducted in the Netherlands is limited. Furthermore, existing research is often outdated and consumer preferences for environmentally friendly products have changed drastically over the last few years. In addition, the corona crisis has given an extra boost to the demand for organic products (Ecovia Intelligence, 2020). This demand is expected to remain strong, even after the pandemic. According to Ecovia Intelligence (2020), previous food and health scares caused spikes in sales followed by sustained demand for organic products.

In the next section, existing studies on purchase decisions for organic food products are discussed, after which several hypotheses are formulated. In Section 3, the different data sets and their limitations are discussed. Section 4 introduces the methods used in this study and in Section 5, these methods are applied to the different data sets and the results are examined. Finally in Section 6, the most important results are summarized and some crucial limitations are discussed. Furthermore, recommendations for future research are made.

2 Theoretical framework

2.1 Goals of ecolabeling

The goal of ecolabeling is to create market-based incentives for environmentally friendly behaviour (Roheim et al., 2011). In a report focused on the environmental impact of food production and consumption in the United-Kingdom, Foster et al. (2007) find that meat, poultry and related products account for around 12% of the global warming potential in the European Union, and milk and dairy products account for approximately 5%. For fruits and vegetables, this percentage is approximately 2%. Ecolabels provide the consumer with information about the environmental impact of products that would otherwise be unobservable. If consumers value the presence of ecolabels and prefer these products over products without ecolabels, this could give room for a price premium that introduces a market incentive for producers to supply these consumer goods (Roheim et al., 2011).

2.2 Predicting organic consumption behaviour

2.2.1 Factors affecting organic consumption

Hemmelskamp & Brockmann (1997) present a list of factors that affect the choice for "green" consumption. These factors are consumer satisfaction, social values, personal utilitarian values, effectiveness, costs, and availability. The study points out that environmentally conscious behaviour could conflict with consumer satisfaction when it does not coincide with other personal needs. Galarraga Gallastegui (2002) points out that many environmentally friendly products may not meet important consumer criteria such as price, performance and quality. Social values are more likely to result in environmentally friendly actions if this goes hand in hand with prestige and recognition (Hemmelskamp & Brockmann, 1997). Smith (2014) also points out the effects of social pressure on consumption. If consumption behaviour would be explained in terms of relative effects, meaning that consumers strive to be greener than others, visibility of consumption and purchase could be considered an important motivator for buying organic products (Galarraga Gallastegui, 2002).

The third important factor Hemmelskamp & Brockmann (1997) describe is identification. If the consumer believes that organic food products fit their lifestyle and identity, this could result in environmentally conscious behaviour. Based on these findings, the first hypothesis is formulated as follows:

H1: *Social values and norms are important determinants of organic consumption behaviour.*

Shafie & Rennie (2012) find that not only product attributes such as nutritive value, taste, freshness and appearance influence consumer preferences towards organic food products, but also concerns about food safety, health and environmental concerns. Padel & Foster (2005) support this by pointing out that personal health remains one of the strongest motivators for organic food consumption. Altruistic motives, such as environmental concerns and animal well-being, are also becoming more important (Padel & Foster, 2005). Paul & Rana (2012) found that the main reasons for purchasing organic food products are related to health concerns. However, the overall benefit of organic products (e.g., environmentally friendly and healthier) is what consumers seek in organic food. Following these findings, the second hypothesis is introduced:

H2: *Personal reasons and motives are important determinants of organic consumption.*

Padel & Foster (2005) find, based on a survey held in the United Kingdom, that consumers that regularly consume organic food products are more likely to be highly educated, of higher social class and more affluent. Furthermore, frequent organic shoppers are often older and have fewer children. A different survey found that most consumers of organic products are often between 45 and 54 years old. A general trend is found that the youngest age groups are the least concerned with environmentally friendly consumption. Most studies conducted on green consumption behaviour find that organic buyers are mostly women (Krystallis et al., 2006a). The study claims that age does not play an important role but that younger people seem slightly more willing to buy green products, thus contradicting Padel & Foster (2005). Krystallis et al. (2006a) also claim that disposable income mainly affects the quantity of organic products bought and not the willingness to buy in general. Shafie & Rennie (2012) find that price premiums on organic food products continue to slow down green consumption. Organic products are usually seen as the expensive alternative and practically inaccessible to lower class households. Based on previous findings, the third hypothesis is formulated as follows:

H3: *Income and age are important determinants of organic consumption.*

Paul & Rana (2012) studied the effects of different demographic factors on the preference for organic food products. They found that education positively affects the quantity of organic food products purchased. Higher-educated consumers were found to be more likely to purchase organic food products. Other demographic factors such as gender were not found to have a statistically significant effect. Based on these findings, the fourth hypothesis is formulated:

H4: *Educational level is an important determinant of organic consumption.*

2.3 Identifying WTP for organic food products

2.3.1 The organic price premium and WTP for organic food products

Due to increasing demand for organic food products and products with ecolabels, supply and therefore the competition for this type of product has increased (Galarraga Gallastegui, 2002). Klein & Leffler (1981) argue that the price of products can be used as an indicator for product quality. Therefore, the price premium on organic products can be regarded as proof of higher product quality, which not only positively affects the demand side for this type of product, but also the supply side (Galarraga Gallastegui, 2002). The literature consistently finds significant price premiums on sustainable products. Roheim et al. (2011) find a statistically significant premium of 14.2% on ecolabeled seafood products in the United Kingdom. Peattie (1995) finds that more than 25% of the British population is willing to pay a premium of up to 25%

for products with an improved environmental performance. In Canada, a public survey found that 64% of the respondents is willing to pay a price premium of 10% for an ecolabeled product (Galarraga Gallastegui, 2002). A 5% price premium was found in Singapore (Galarraga Gallastegui, 2002; Jha et al., 1993) and a 5 to 7% premium in the United States (Van Ravenswaay & Hoehn, 1991). Aschemann-Witzel & Zielke (2017) study research on consumer behaviour towards the price of organic food from 2000 to 2014. They find that on average, the willingness to pay a premium for more sustainable food alternatives is 30%. However, they find strong differences between different studies and different types of organic food products. According to Padel & Foster (2005), the main reason for not buying organic food remains, however, price. Aschemann-Witzel & Zielke (2017) establish that price knowledge and price sensitivity among organic consumers are relatively low in comparison to non-organic consumers. Overall, different price premiums are found in different countries which points out that the current research on WTP for ecolabels and their implications are not one-size-fits-all. Given these findings, the fifth hypothesis is formulated:

H5: *The presence of an ecolabel positively affects purchase intentions and WTP for food products.*

According to Laroche et al. (2001), consumer attitudes towards the environment are very good predictors of WTP for organic products. A Spanish study showed that only consumers who normally consume organic food and are concerned about their health were willing to pay a price premium of 15 to 25% for organic food products (Gil et al., 2000). Krystallis & Chryssohoidis (2005) find that WTP differs for different types of products in Greece. They find that Greek consumers exhibit the highest WTP for organic fruits and vegetables. Given these findings, the sixth hypothesis is formulated as follows:

H6: *Purchase intentions and WTP for organic food products differ for different types of products.*

Given that this study focuses on Dutch consumers, it is important to establish what literature currently exists on the Dutch market and what the results of the existing studies may imply for this research. Bunte et al. (2010) performed a real-life experiment to determine the price-sensitivity of demand for organic products. Based on supermarket scanner data, they find that there is low price elasticity of demand for organic products as well as low WTP for the social attributes of these products based on their findings for ten Dutch communities in 2006. However, the research excluded high-income communities in order to represent the average Dutch community. Furthermore, large communities and very small communities were excluded as well due to organizational constraints.

3 Data

3.1 Primary survey data

Data are collected with the help of a survey that is divided in four sections: the first section consists of questions related to respondent demographics. In this section, respondents are asked about their gender, age, education, employment status, income, and household size. The second section measures the respondent's purchase intention for organic food products. In this section, the respondent is randomly shown either a food product with an ecolabel or the same food product without an ecolabel. For each product, the respondent has to indicate how likely he or she is to buy this product on a 5-point Likert scale, which ranges from "extremely unlikely" (value 1) to "extremely likely" (value 5). To minimize the effects of personal preferences on purchase likelihood for certain food products, the respondent is given the option to indicate that he or she would never buy a product due to personal taste and/or preferences.

The third section measures WTP for organic food products in comparison to non-organic food products. This is done with the help of a simple conjoint analysis in which the respondent is shown two options of the same food product of which the presence of an ecolabel and prices differ. The respondent is shown the organic food product with a price premium that varies per product, but is always higher than the price of the non-organic food product. In specific, the respondent is shown 5 different food products: bananas, tomatoes, milk, chicken and spaghetti. All five products are shown 3 times: once with the actual supermarket prices, once with a price increase of 10% for the organic alternative relative to the original supermarket price, and once with a price decrease of 10% for the organic alternative relative to the original supermarket price. The price of the non-organic product remains equal to the supermarket price for all three choice questions. This format allows us to not only measure the change in demand for organic products after an increase in price, but also the change in demand if the prices for these products would be lowered. Examples of the survey questions from section two (measuring purchase likelihoods) and three (conjoint analysis) can be found in Appendix A.

The survey also considers the effects of personal preferences that are only focused on certain types of food products. Because of the relatively higher environmentally impact of meat products, some may prefer the organic option for these types of products, but not for other types of products such as fruits and vegetables. Padel & Foster (2005) also point out the variation in purchase likelihood between different product categories in the decision-making process for organic food. Therefore, each section shows a balance of different types of food products. In both sections, the respondents are shown milk, chicken, bananas, tomatoes and spaghetti. This allows us to measure whether a price premium on organic food products may only be tolerated for food products that are considered relatively environmentally unfriendly. Furthermore, all previously mentioned products belong to the top 10 most frequently bought products in their product category according to The Netherlands Nutrition Centre.

The fourth and final section of the survey checks whether the respondent's choices in the previous sections coincide with his/her beliefs on their behaviour towards organic food products. The first question asks about their awareness of ecolabels on food products. The second question measures whether the respondent believes that his/her consumer choices are affected by the presence of ecolabels. The final question is about the main reason for choosing an organic product over a non-organic product. The final three survey questions and their answer options can be found in Appendix A.

To collect responses, the survey was distributed among various social media platforms, as well as posted on several respondent-collection websites. Furthermore, the survey was distributed in the area of Rotterdam by mail, to ensure a more even distribution of age and gender in the sample. This resulted in 360 respondents, of which 101 did not fully complete the survey. From all 257 respondents who completed the survey, 39% indicated to be male and 59% female, the other 2% preferred not to say.

Figure 1 shows the distribution of age and income groups within the sample. It can be seen that the majority of the sample is between 18 and 24 years old. The age groups 25-34 years, 45-54 years, and 55-64 years, are all more or less the same size. The smallest age group in the sample is that between 12-17 years (0.77%), which makes sense given that this group is normally not faced with these types of purchasing decisions.

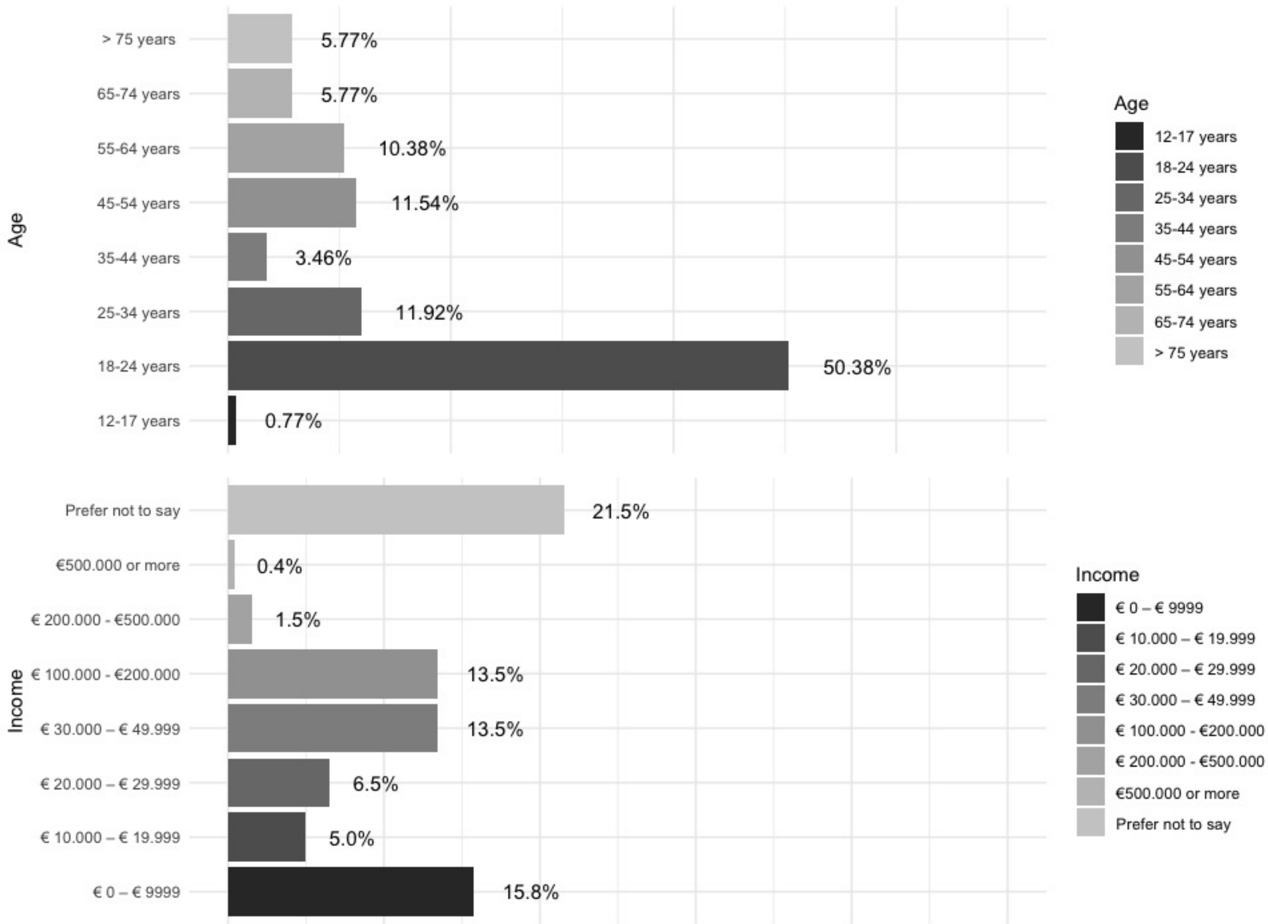


Figure 1: Distribution of age and income for survey data.

Figure 1 also shows that the largest part of the sample prefers not to reveal their income. From those who have stated their income, most respondents fall in the lowest income category (€0 - €9,999). The second largest income groups are the income groups with earnings between €30,000 and €49,999, and €100,000 and €200,000 per year.

Figure 2 shows the distribution for the variables *Education* and *Household size*. It can be seen that the highest proportion of respondents corresponds to those having completed a Bachelor's degree and the second highest proportion to respondents having completed a Master's degree as their highest level of education. This shows that the sample is relatively highly-educated. The figure also shows that most respondents live in a two-person household, followed by a four-person household.

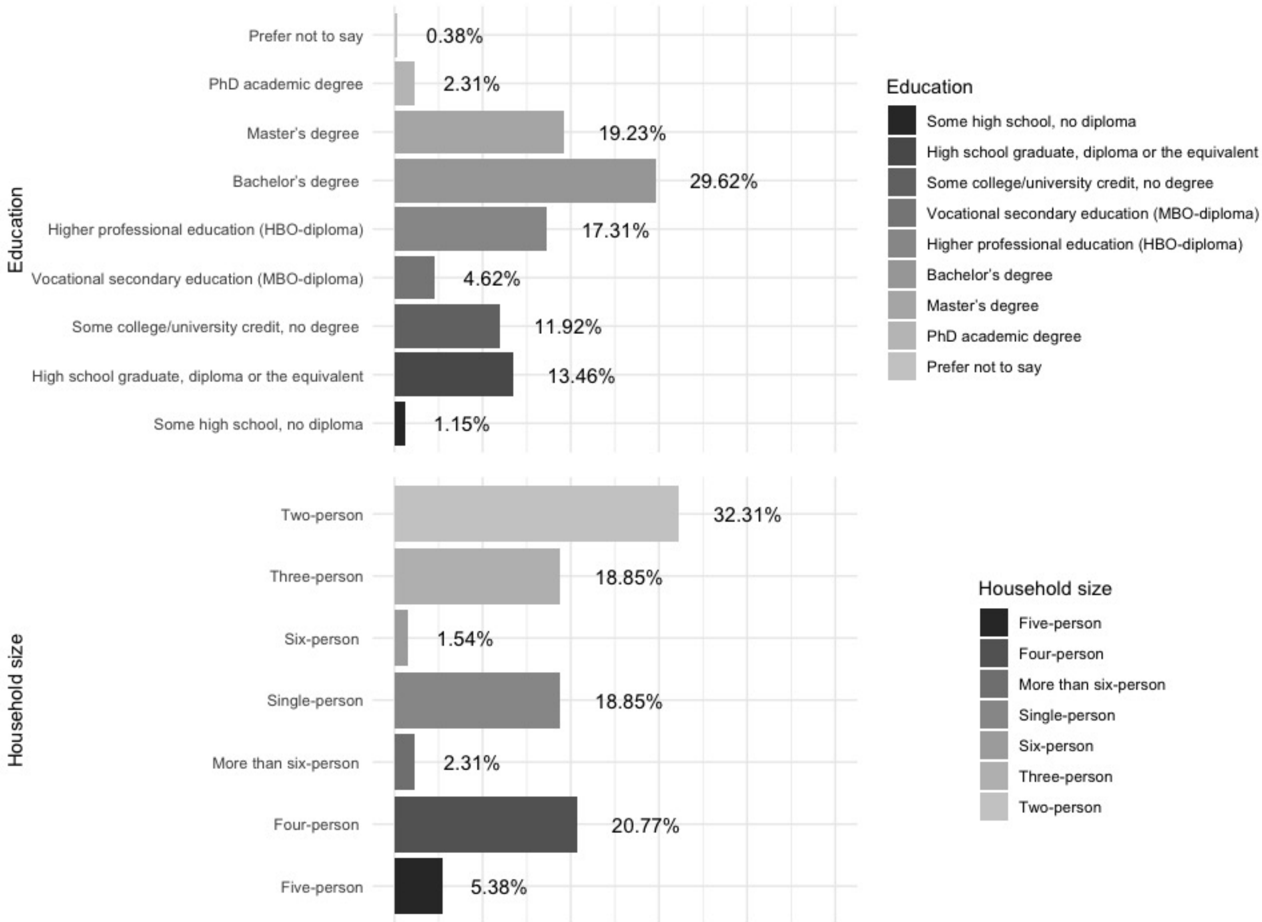


Figure 2: Distribution of education and household size for survey data.

Figure 3 shows the distribution of employment among the respondents. It can be seen that more than 40% of the respondents is student, more than 35% of the sample is employed for wages, followed by a little over 10% that is retired. Only a very small part of the population reports to be unemployed or unable to work.

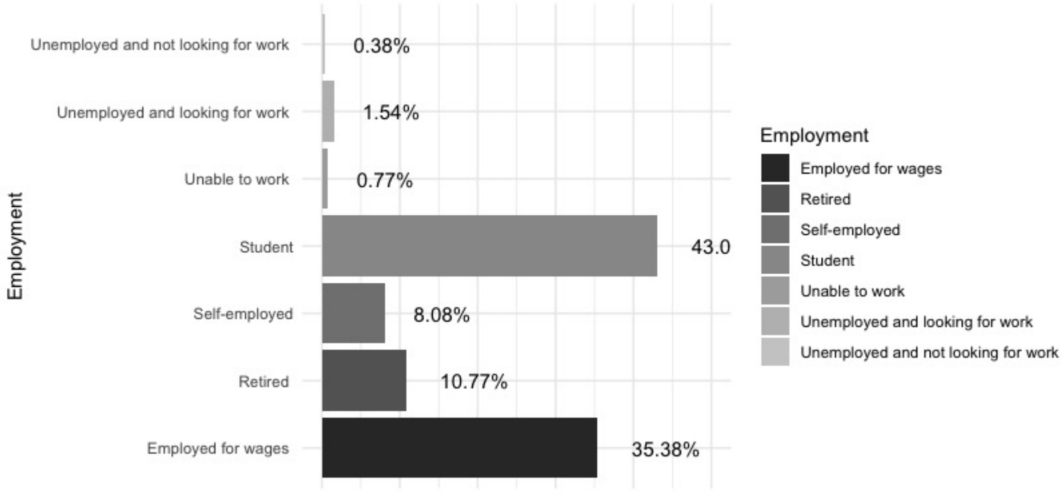


Figure 3: Distribution of employment in the sample.

Figure 4 shows the distribution of purchase likelihood for the five products and their organic and non-organic variants shown in the second section of the survey. The Figure shows that the differences in purchase likelihood are relatively small for most food products. In specific, the differences in distributions of purchase likelihood for bananas and tomatoes are almost negligible. For chicken, the non-organic variant has a slightly higher purchase likelihood than the organic variant.

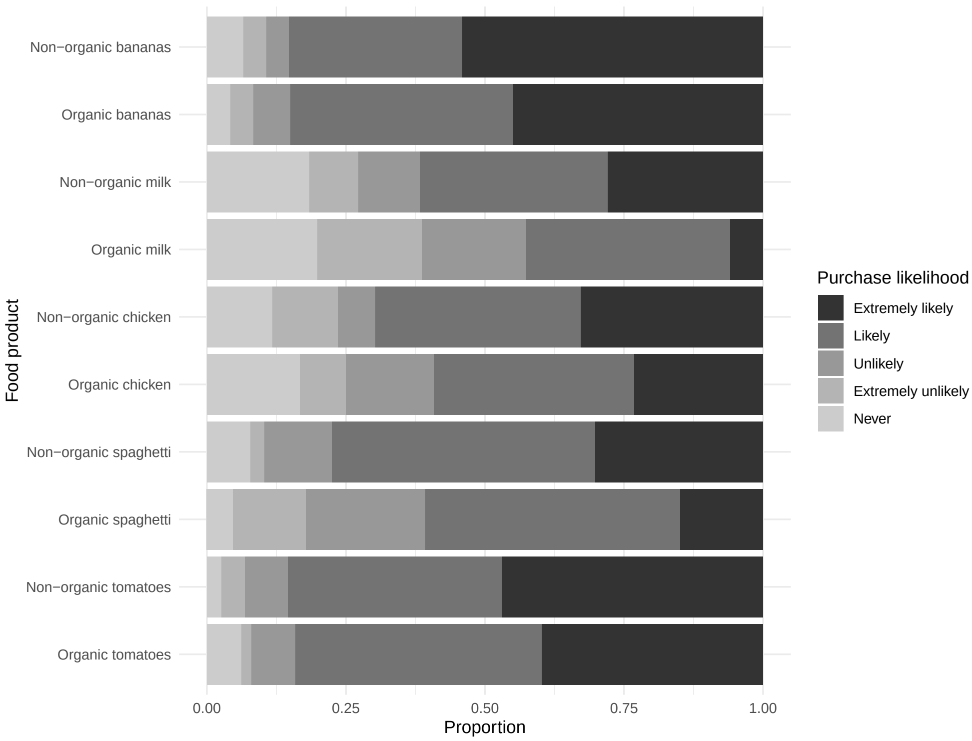


Figure 4: Distribution of purchase likelihood for all organic and non-organic variants of the five food products (chicken, spaghetti, tomatoes, milk and bananas).

The products that show the largest difference in purchase likelihood between its organic and non-organic variant (based on cumulative proportions of positive likelihoods) are milk and spaghetti. A smaller, but visible difference can be seen for chicken. In Section 5, the significance of these differences are tested with Mann-Whitney U tests.

The third part of the survey, which involves consumer choices between organic and non-organic alternatives, allows us to divide consumers into one of four different classes:

1. Consumers that will never buy the organic product
2. Consumers that will only buy the organic alternative when the price is 10% lower than the original supermarket price
3. Consumers that will only buy the organic alternative when the price is equal to the original supermarket price or 10% lower
4. Consumers that will always buy the organic alternative, even with a price increase of 10%

For every product, the distribution between the different consumer types is visualized in Figure 5. It can be seen that for all products, the majority of the respondents never opts for the organic alternative if they have to choose between the non-organic (more affordable) variant and the organic (more expensive) variant. The product that has the largest group of respondents only preferring the organic alternative (consumer type 4) is spaghetti. Intuitively, this is in contrast with the results shown in Figure 4 that relate to the first part of the survey, where the difference in purchase likelihood for the organic and non-organic variant of spaghetti were shown to be the largest. On the other hand, the results are not as surprising given that the price difference between the organic and non-organic alternative for spaghetti is relatively small and the product itself is in general, relatively inexpensive. Figure 5 also shows that for spaghetti and tomatoes, none of the respondents fall under consumer type 2: consumers only choosing the organic alternative when the price is 10% lower than the supermarket price. The product which incentivizes most respondents to switch to the organic alternative are bananas, with around 13% of all respondents choosing the organic alternative only when the price is lower than the supermarket price.

Figure 5 also introduces a significant limitation of the survey data, namely the lack of variation in consumer types and the very small proportion of respondents that react to a price change. The majority of the respondents either always opts for the organic alternative or never opts for the organic alternative, and only a very small part of the consumers show to be prone to switching when the price of the organic variant changes. This already provides us with strong evidence that the mixed logit model will most likely not yield very informative results, and that the only data that may be useful for this analysis are the data relating to bananas, since this data show the most variation of all food products. Therefore, to support the findings of the mixed logit model, data from the Liss Panel Archive are used. This data set is discussed in Section 3.2.

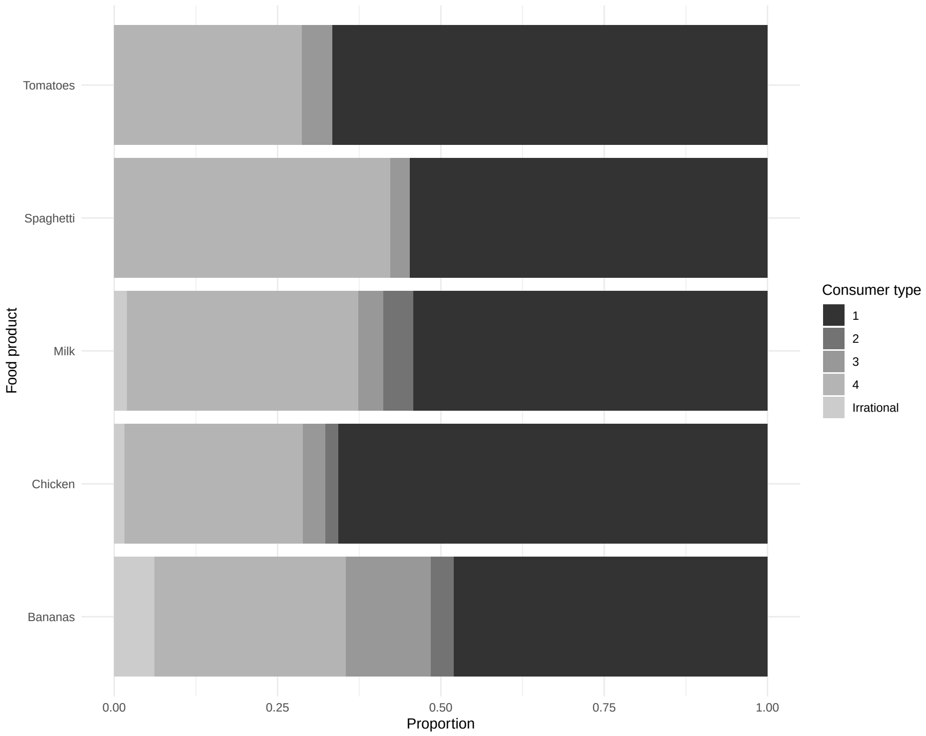


Figure 5: Distribution of consumer types for all five food products: chicken, spaghetti, tomatoes, milk and bananas.

In the final part of the survey, respondents were asked three questions to gain information about (1) their awareness of ecolabels on food products, (2) their view on how this presence affects their purchasing behaviour and (3) what their main reason would be for purchasing organic food products. Figure 6 shows that most respondents indicate to be aware of ecolabels on food products. Around 19% of the sample indicates to be either unaware or very unaware of these labels. When looking at the middle plot, it is evident that most people agree with the statement that the presence of ecolabels affects their purchase behaviour. However, a relatively great part of the respondents (around 29%) indicates that they neither agree nor disagree with this statement. Around 15% of the respondents believe that the presence of ecolabels does not affect their purchase behaviour. The lower plot shows that around 38% of the respondents buy organic food products to be more environmentally friendly, followed by 24% choosing organic because they believe it to be more socially acceptable. Less than 10% indicates to buy organic food products because it is tastier and 5% indicates to never buy organic food products. From the respondents who answered “other”, some mentioned that they prefer buying organic if the packaging does not involve plastic, or if the product does not involve the use of pesticides and dangerous chemicals. Others mention that their main reason for buying organic often differs and that for some products they simply find the organic variants tastier, while for other products it’s because they believe it to be more animal friendly.

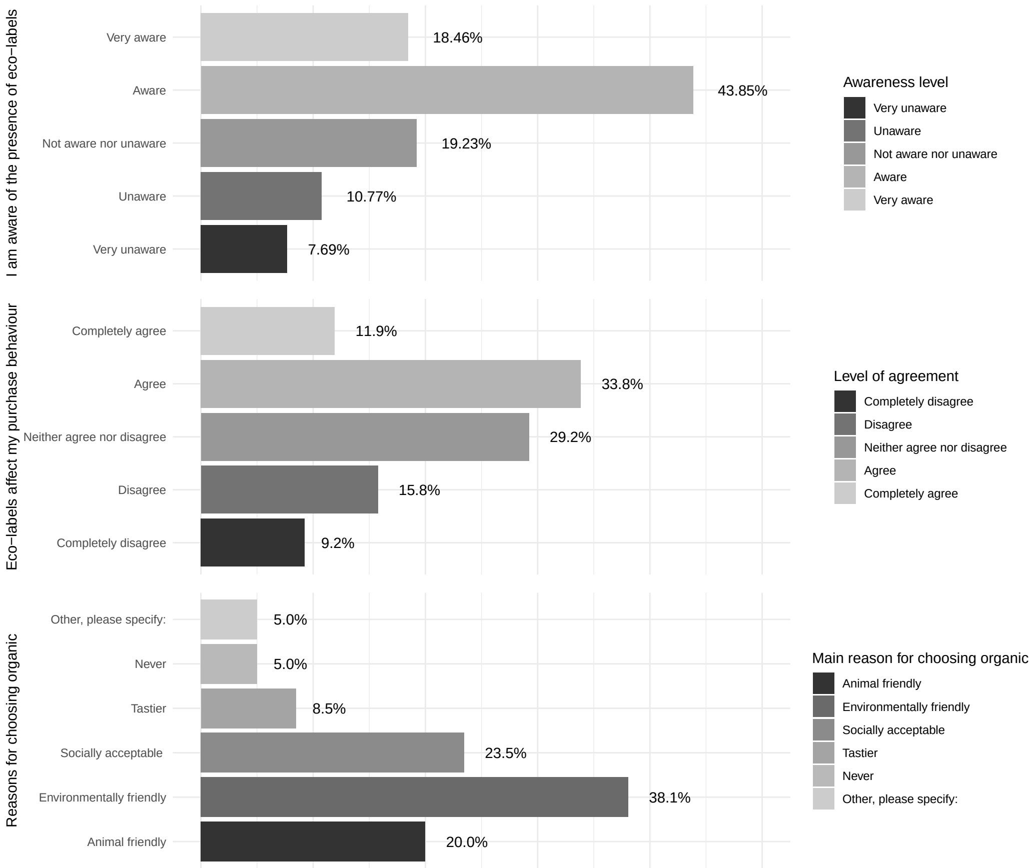


Figure 6: Distributions of answers for final 3 questions of survey data

3.2 Liss Panel data

In the second part of the analysis, WTP for organic food products is measured with the help of mixed logit models. As discussed in Section 3.1, Figure 5 provides strong evidence that the survey data are not suitable for the model proposed, and supporting data are necessary to draw reliable conclusions from the results. Therefore, data from the Liss Panel Archive on consumer heterogeneity with respect to morality in consumption decisions and perceptions of animal welfare are used. This data set contains survey data collected from 1700 respondents in 2012. In the survey, respondents were asked several questions about animal welfare and shown different meat products with different prices and packaging. Similar to the survey created for this study, respondents were asked to choose between two different food products. Respondents were shown different choice questions for the same product, which was either chicken breast or pork cutlet. This study only focuses on the chicken breast data, as this product is also used in the conjoint analysis of the primary survey data discussed in Section 3.1. For this specific product, consumers were shown 18 different choice sets and the order of the choice sets were randomized. Furthermore, the respondents were shown different choice sets based on a specific version. The survey consisted of a total of 28 different pictures of chicken breast, divided under 18 versions. Thus, consumers having the same version were shown the same choice sets, but in a different (randomized) order.

The pictures the respondents were shown differed based on price per kilogram, ecolabel, the number of stars if an ecolabel was present, a label with the name of the product (“kipfilet”, meaning chicken breast in Dutch), and a third label with information on the product. This third label could contain different messages about the origin of the product, the flavour of the product, an explanation relating to the number of stars on the ecolabel, or something completely random such as "just act normal, then you're already crazy enough". An example question of the survey can be found in Figure 7.

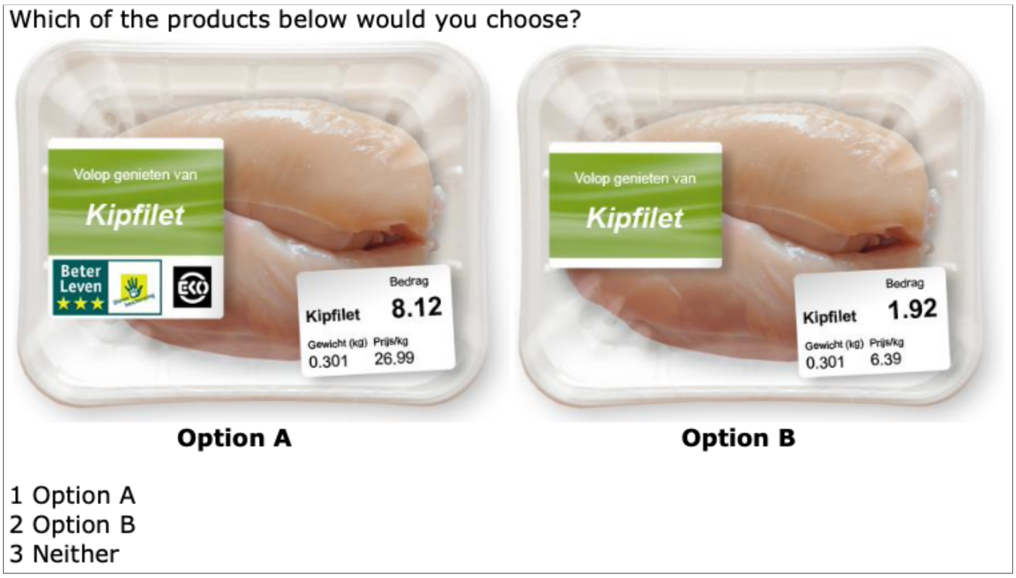


Figure 7: Example of choice set presented in Liss Panel study.

The total number of choice questions across all individuals for the chicken breast data is 12,522 after the removal of missing data points. From this number, 5,926 choice questions involved a product with an ecolabel versus a product without an ecolabel. Focusing on these choice questions allows us to analyze the variation in the Liss Panel data regarding preferences for organic products. From all choice questions involving organic versus non-organic chicken breast, 70% corresponds to the respondent preferring the cheaper, non-organic variant over the more expensive, organic variant.

To further analyze the variation in the data, we can look at the percentage of respondents choosing the non-organic chicken breast, when presented with the choice question involving the lowest price difference: the non-organic chicken breast for €2.25 and the organic chicken breast for €2.71. In this case, 56% of the respondents prefers the non-organic variant, and 44% the organic variant. When having to choose between non-organic chicken for €1.50 and organic chicken for €2.71, these percentages become 60% and 40% respectively. When faced with a higher priced organic chicken (€3.16), and the €2.25 non-organic variant, these percentages remain the same. The percentage of individuals preferring the non-organic alternative increases to 65% when the price for organic chicken increases to €3.73 and to 80% when the price jumps to €5.72. As the only two prices that respondents are shown for non-organic chicken are €1.50 and €2.25, the biggest changes in preferences are likely to occur when facing the choice question between organic and non-organic chicken, with the price of organic chicken rising from €3.72 to €5.72. Focusing on these choice questions may therefore lead to more informative results in comparison to when all choice questions are included in the analysis.

4 Methodology

4.1 The ordered logit

In the first part of the research, an ordered logistic regression is performed for every food product based on four consumer classes that can be derived from the data. As previously stated, these consumer types are:

1. Consumers that will never buy the organic product
2. Consumers that will only buy the organic alternative when the price is 10% lower than the original supermarket price
3. Consumers that will only buy the organic alternative when the price is equal to the original supermarket price or lower
4. Consumers that will always buy the organic alternative, even with a price increase of 10%

Given that the consumer groups are in a natural order from least "organic-oriented" to most "organic-oriented", based on their preferences for organic products, an ordered logit can shine light on the probability of falling into a specific consumer class given the individual characteristics of the consumer. Furthermore, the outcomes of this model could provide valuable information about the consumer characteristics that are most important in determining purchase behaviour in relation to organic food products.

For the ordered logit, Y is the response variable presenting an ordinal outcome with j categories, and x presents the covariates in the model. Then $P(Y \leq j)$ is the cumulative probability that Y is smaller than j (Bruin, 2021). The odds of Y being smaller or equal to a particular category j is then

$$\frac{P(Y \leq j)}{P(Y > j)}. \tag{1}$$

Note that this is only the case for $j = 1, ..., J–1$ since $P(Y > j)$ for the highest category would be 0.

The following derivations follow from McCullagh (1980). If the odds of $Y \leq j$, given covariates x would be given by $\kappa_j(x)$, the proportional odds model under the condition that $j = 1, ..., J–1$ can be written as follows:

$$\kappa_j(x) = \kappa_j exp(-\beta^T x).$$

Here, β is a vector of unknown parameters. The ratio of the corresponding odds is then given by

$$\frac{\kappa_j(x_1)}{\kappa_j(x_2)} = exp(\beta^T (x_2 - x_1)).$$

This ratio is independent of the category j and solely depends on the difference between the values of x_1 and x_2. Since the odds for the event are equal to the ratio given in equation (1), the proportional odds model, a linear logistic model can be obtained by taking the log of the ratio, which results in equation (2),

$$log \frac{P(Y \leq j)}{P(Y > j)} = \theta_j - \beta_x^T. \tag{2}$$

Here, $\theta_j = log(\kappa_j)$ so that the model fulfills the proportional odds assumption stating that the difference between the cumulative logits is independent of the category (McCullagh, 1980).

In the survey context we estimate the following model for all five food products:

$$
\begin{aligned}
logit(P(Y \leq j)) = \{&\theta_j - \beta_1 * gender - \beta_2 * age - \beta_3 * income \\
&- \beta_4 * education - \beta_5 * household\ size - \beta_6 * employment\}.
\end{aligned}
$$

The parameters are obtained with maximum likelihood estimation. For the model estimation, the `polr()` function from the `MASS` package in R is used.

4.2 Predicting consumer purchase behaviour

The first research question aims to predict organic consumption behaviour. Consumers are classified into organic and non-organic consumers based on their answers to the first set of choice questions relating to the choice between the organic and non-organic food products at their current supermarket prices. If the respondent chose the organic alternative against the actual supermarket price, he or she is thus classified as an organic consumer. The prediction data therefore disregard the consumer's purchase decisions for the fictional prices that were shown afterwards. For the prediction, two machine learning methods are used and compared against each other to see which method performs best. These methods are support vector machines (SVM) and random forests. Both methods are discussed in detail below.

4.2.1 Random Forests

Random forests is a machine learning method based on a combination of tree predictors. It is an ensemble learning method that can perform both regression and classification tasks. In this study, the model is used to classify respondents into organic and non-organic consumers. The model does so by using an ensemble of decision trees. The structure of a single decision tree is visualized in Figure 8. The Figure shows that a single decision tree begins at the root node. Based on the input of the model, the information is passed through to the child nodes, or decision nodes. A leaf node marks the end of a decision tree and corresponds to a class of the output variable. In other words, the leaf nodes indicate whether a respondent is classified as an organic or non-organic consumer. The independent variables serve as the input of the model and determine where the tree splits. For example, the tree could first split based on age, where respondents below the age of 30 would immediately be classified as organic consumers. Re-

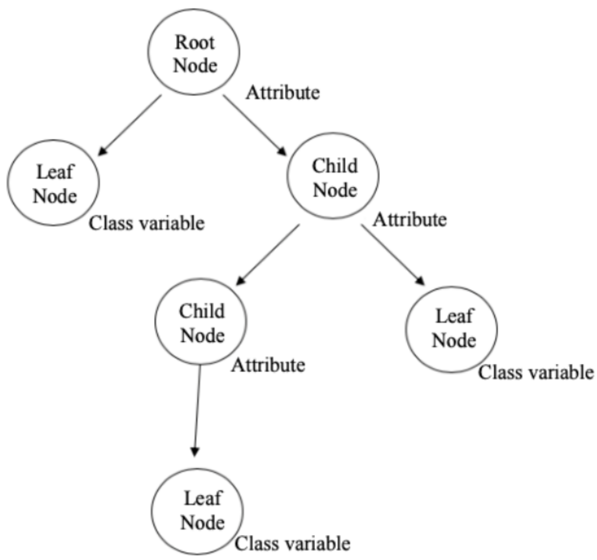


Figure 8: Decision tree structure (Martínez et al., 2009)

spondents above the age of 30 would then be split again based on a second variable, for example income, where respondents with an income below €30,000 per year would be classified as non-organic consumers. The resulting non-classified consumers would then again be split based on a third variable, and this process then continues until all respondents are classified.

Breiman (1996) demonstrates that substantial gains in classification accuracy can be obtained by using ensembles of decision trees. Each tree in the collection is based on randomly selected input variables and final predictions are obtained by aggregating over the ensemble of trees. In other words, the final classification of an observation is determined by the majority vote of all predictions resulting from the individual decision trees. The outcome of the random feature selection can be visualized by the Gini importance (Menze et al., 2009). At each node of a decision tree, the optimal split is made based on the Gini impurity, which measures how well a split divides the sample into two classes (Menze et al., 2009). For each split, the largest possible decrease in Gini impurity is desired. Accumulating all the decreases in Gini impurity over all nodes for all trees in the random forest, results in the Gini Index. The Gini Index formula can be written as follows:

$$I_G = \sum \sum_{j \neq i} \frac{f(C_i, T)}{|T|} \frac{f(C_i, T)}{|T|}. \tag{3}$$

In equation (3), $\frac{f(C_i,T)}{|T|}$ corresponds to the probability that a particular case in training set T falls in class C_i. The Gini Index indicates how often a particular feature is used to determine the optimal split. It also indicates the discriminative power of a feature for the classification problem (Menze et al., 2009).

An important advantage of random forests is that it can handle a large number of input variables without overfitting. The method is therefore generally considered to be one of the most accurate machine learning techniques (Biau, 2012). In the model introduced by (Breiman, 2001), the best split is calculated within the subset of randomly chosen input variables and no pruning is performed meaning that all trees are grown maximally (Genuer et al., 2008). The random parameter selection prevents the trees from being highly correlated, where the correlation is defined as the ordinary correlation of predictions on so-called out-of-bag (OOB) samples. The OOB sample defines the set of observations not used for building a corresponding tree. It is used to estimate the prediction error and evaluate variable importance (Genuer et al., 2008).

The output of Random Forests exists of the OOB-error rate, the number of trees used for the classification and the number of variables tried at each split. The final model is the random forest with the number of trees and variables that result in the highest prediction accuracy on the training data. In R, the `caret` package is used for parameter tuning and to train the model. Leave-one-out cross validation is used to determine the optimal number of trees and number of variables for a single tree. This method is preferred due to the small number of observations in the data set and the reliable performance estimates it produces.

4.2.2 Support Vector Machines

The following methodology follows from Gunn (1998). SVM are based on Structural Risk Minimization (SRM), which minimizes an upper bound on the expected risk instead of minimizing the error on the training data. This principle causes SVM to have a greater ability to generalize. The goal is to separate different classes by a function obtained from the existing data and to produce a classifier which is able to do this for new, unseen data. The classes can be separated with a linear classifier called the optimal separating hyperplane that maximizes the margin, e.g., the distance between the hyperplane and the nearest data point of each class. The hyperplane is given by

$$\langle w, x\rangle + b = 0.$$

When considering a canonical hyperplane in which the parameters are scaled, the parameters w and b are constrained by

$$\min_i |\langle w, x^i\rangle + b| = 1.$$

According to Gunn (1998), this constraint states that "the norm of the weight vector should be equal to the inverse of the distance, of the nearest point in the data set to the hyperplane." The margin of the optimal hyperplane that we want to maximize is given by:

$$\rho = \frac{2}{\| w \|}.$$

Here, ρ is subject to the following constraints:

$$f(x_i) = \begin{cases} 1, & \text{if } \langle w, x_i\rangle + b \geq 1, \\ -1, & \text{if } \langle w, x_i\rangle + b \leq 1, \end{cases} \tag{4}$$

$$\gamma_i(\langle w, x_i\rangle + b) \geq 1. \tag{5}$$

The hyperplane that optimally separates the data minimizes

$$\phi(w) = \frac{1}{2} \| w \|^2 . \tag{6}$$

This formula is independent of b given that equations (4) and (5) are satisfied. This is because when the value of b changes, the hyperplane will move itself in the normal direction (Gunn, 1998). Equation (6) can be solved with the Lagrange multiplier method. Here, Lagrange multipliers are introduced $\lambda_i \geq 0$ and the following equation is minimized:

$$\phi(w, b, \lambda) = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{N} \lambda_i(y_i(\langle w, x_i\rangle + b)) - 1. \tag{7}$$

Here, N is the number of output vectors. The Lagrangian has to be minimized with respect to w and

b, and maximized with respect to λ. Equation (7) can be transformed in dual form, which simplifies the quadratic optimization problem for λ_i. This dual problem can be obtained by first taking the minimum with respect to w and b and rewriting the equation. This results in the following equation that should be maximized:

$$\max L(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} (\lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle). \tag{8}$$

Equation (8) is subject to the following two constraints:

$$\lambda_i \geq 0, \qquad \sum_{j=i}^{N} \lambda_j y_j = 0.$$

Solving equation (8) under these two gives the optimal hyperplane

$$w^* = \sum_{j=1}^{N} \lambda_i y_j x_i, \qquad b^* = -\frac{1}{2} \langle w^*, x_r, x_s \rangle.$$

Here, x_r and x_s are any support vector from each class satisfying the following conditions:

$$\lambda_r, \lambda_s > 0, \qquad y_r = -1, y_s = 1.$$

This is referred to as a hard classifier. However, a soft classifier may be more appropriate than the hard classifier. This is because when the classifier is probed within the margin, the classifier produces an output between -1 and 1. Given the following Karush-Kuhn-Tucker (KKT) conditions,

$$\lambda_i \leq 0, \qquad \lambda_i (y_i (\langle w, x_i \rangle + b - 1)) = 0,$$

it follows that the only points that satisfy,

$$y_i (\langle w, x_i \rangle + b) = 1,$$

have non-zero Lagrange multipliers. These points are called the support vectors. The number of support vectors can be very small when the data are linearly separable. The optimal hyperplane is thus only determined by a small subset of training data. However, not all data are linearly separable. In this case, slack variables are introduced which can be regarded as a measure of the misclassification errors. These slack variables satisfy the non-negativity assumption $\xi \geq 0$. Now, a penalty term is added and the following equation is minimized to obtain the optimal hyperplane

$$\phi(w, \xi) = \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{N} \xi_i^k. \tag{9}$$

Here, C is a given value. A large value of C favors overfitting at the cost of a small margin, while a small C accounts for small errors and the margin is allowed to be large. To determine the optimal value of the cost parameter C as introduced in equation (9), cross-validation will be used. Equation (9) is subject to the following conditions:

$$f(x_i) = \begin{cases} 1, & \text{if } \langle w, x_i \rangle + b \geq 1 - \xi_i, \\ -1, & \text{if } \langle w, x_i \rangle + b \leq 1 + \xi_i, \end{cases}$$

which is also called the soft margin approach.

When a linear boundary is inappropriate, the data should be transformed into a higher dimensional space. Beforehand, a non-linear mapping is chosen and the SVM constructs the optimal hyperplane in the higher-dimensional space. The data can always be separated in a higher dimensional space with the help of a transformation function. The original space does not have to be transformed, only the appropriate Kernel function has to be chosen. Among these Kernel functions are the linear function, the polynomial function, the Gaussian Radial Basis Function, and the Sigmoidal Function. In this research, several Kernel functions are used and compared against each other. The dual Lagrangian optimization problem then becomes:

$$\max L(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j K(x_i, x_j).$$

Here, K is the kernel function performing the non-linear mapping into feature space (Gunn, 1998). The constraints remain unchanged. In order to determine the appropriate kernel function, a hyperparameter search can be set up and different kernel functions can be compared against each other. Based on the accuracy performance measure, the appropriate kernel function for the prediction task is determined.

Similar to the random forest model, the `caret` package in R is used with leave-one-out cross validation to tune the hyperparameters and obtain the final model.

4.3 The Mixed Logit model

In the second part of the analysis, a mixed logit model is used to model the probability of a consumer choosing an organic (or non-organic) alternative, based on the attributes of the product. The attributes used in this study are price and ecolabel. The mixed logit differs from the traditional logit framework in that it allows for consumer heterogeneity by estimating distributional parameters instead of fixed parameters that are assumed to be the same for all individuals. The mixed logit model also allows us to model the distribution of WTP.

According to McFadden & Train (2000), the mixed logit allows for approximation of any discrete random utility choice model, if specified properly. It overcomes the limitations of the standard logit by allowing for correlations over time and random taste variation (Train, 2009). Since the mixed logit assumes that the parameters of the model vary across individuals, it can take the heterogeneity of the population into account (The Comprehensive R Archive Network, 2021). In order to understand the mixed logit, first consider the standard logit model. The model equation would then be written as follows:

$$Log(\frac{P(Choice = 1)}{1 - P(Choice = 1)}) = \beta_0 + \beta_1 Price + \beta_2 Ecolabel + \varepsilon.$$

Here, the probability of choosing a certain option is modelled, where the consumer will choose the organic alternative if the probability of choosing this product for a certain price is higher than this probability for the non-organic alternative against a certain (lower) price. Here, the coefficients are assumed to be the same for all consumers. The mixed logit on the other hand, assumes that all beta's differ per individual but follow the same distribution, and can therefore be regarded as random variables drawn from the same (typically normal) distribution. The parameters that the model produces are generated from this distribution via simulation. The probabilities are based on the underlying individual utility function for individual i and alternative j in time t,

$$U_{ijt} = -\alpha_j p_{ijt} + \beta_j x_{ijt} + \varepsilon_{ijt}, \quad j = 1, ..., J; \quad i = 1, ..., N, \tag{10}$$

where p relates to the price attribute and β to the ecolabel attribute. The error term represents all factors that affect the individual's utility but are not captured by the product attributes in the equation. The utility is known to the decision maker but not to the researcher (Train, 2009), and the decision maker chooses the alternative that yields the highest utility. The probability that individual i chooses alternative j over all other alternatives k in time t can then be written as

$$P(Choice = 1) = \frac{exp(-\alpha_j p_{ijt} + \beta_j x_{ijt})}{\sum_{k=1}^{J} exp(-\alpha_k p_{ikt} + \beta_k x_{ikt})}.$$

From the utility function specified in equation (10), WTP can be derived. WTP for an attribute can be derived by taking the ratio of the attribute (ecolabel) coefficient to the monetary (price) coefficient (Hole & Kolstad, 2012). This is the conventional approach and is also referred to as calculating WTP in the preference space. The conventional formulation of utility could, however, result in distributions of WTP that are heavily skewed. Train & Weeks (2005) therefore suggest to use the definition of WTP to rewrite the utility equation such that the coefficients of the model already represent the desired WTP measure. This model is called the WTP space model. The derivations and assumptions of these models are explained in detail in section 4.4.

The parameters of the mixed logit are obtained with maximum likelihood estimation, which is done by minimizing the negative log-likelihood function. In R, the `logitr()` package is used to estimate the models, as this package allows for the derivation of WTP in both the preference space and the WTP space. A more in-depth, technical specification of the mixed logit model can be found in the next section.

4.3.1 Mixed logit technical specification

The following derivation of the mixed logit model follows from Train (2009). The probabilities of the mixed logit can be obtained by taking the integrals of the standard logit probabilities over a density of parameters. Train (2009) therefore notes that every model whose choice probabilities can be expressed in the following form,

$$Z_{ij} = \int P_{ij}(\beta) f(\beta) d\beta, \tag{11}$$

can be regarded as a mixed logit model. Here, $f(\beta)$ is a density function and $P_{ij}(\beta)$ is the logit probability evaluated at parameters β which follow a certain distribution. If utility is linear in β, the mixed logit probability takes its normal form

$$P_{ij} = \int \frac{e^{\beta_i x_{ij}}}{\sum_k e^{\beta_i x_{ik}}} f(\beta) d\beta. \tag{12}$$

This is the unconditional choice probability, derived from the integral of $P_{ij}(\beta)$ over all possible values of β_i. If the density function $f(\beta) = 1$ when $\beta = b$ and 0 otherwise, the standard logit function is obtained

$$P_{ij} = \frac{e^{\beta_i x_{ij}}}{\sum_k e^{\beta_i x_{ik}}},$$

where the probability relates to the probability of individual i choosing alternative j over all other alternatives. As mentioned before, in the mixed logit model, we assume that the values of β_i are random draws from a distribution whose parameters are estimated. The mixed logit probability is therefore a weighted average of the standard logit at different levels of β and the weights are given by the density function $f(\beta)$. A distribution for the parameters can be specified by the researcher. The two most common ones are the normal and log-normal distribution. The latter can be used when the coefficients are expected to have the same sign for all individuals. For example, the price coefficient could be assumed to be negative for all consumers. In the case of organic food products however, this does not necessarily have to be true, as higher prices may reflect higher quality and more environmentally friendly behaviour. Organic consumers may therefore have positive price coefficients if they always prefer the organic alternative.

In the survey context, there are repeated observations for the same individual. Therefore, the longitudinal dimension of the data should be taken into account. In the simplest specification, the coefficients that enter utility vary across people but are constant over choice situations. If we consider a sequence of alternatives, one for each choice situation, z where $z = z_1, ..., z_T$, the conditional probability that the individual makes this sequence of choices is the product of logit formulas:

$$P_{iz}(\beta) = \prod_{t=1}^{T} \frac{e^{\beta_i x_{izt}}}{\sum_k e^{\beta_i x_{ikt}}}.$$

The only difference between the mixed logit with one choice and the one with repeated choices is that the integrand as introduced in equation (11) now contains the product of logit formulas; one for each time period (Train, 2009). Again, a draw of β is taken from its distribution. For each time period, the logit formula is calculated and the products of these logits are taken. This process is then repeated for many draws and the coefficients are obtained by averaging the results, which we call estimation by simulation. The simulation process is explained in detail below.

Given that exact maximum likelihood estimation is not possible, the probabilities can be approximated through simulation. Now, we assume that the coefficients β_i are distributed with density $f(\beta|\theta)$, where θ reflects the parameters of the distribution (such as the mean and covariance of β) (Train, 2009). The choice probabilities now become

$$Z_{ij} = \int P_{ij}(\beta) f(\beta|\theta) d\beta.$$

The simulation process looks as follows: First, draw a value of β from the density function $f(\beta|\theta)$ and label it as the first draw. Then, calculate the logit formula as presented in equation (12), using the value of β obtained from this draw. Finally, repeat these two steps many times and average the results. This final step results in the simulated probability

$$SP_{ij}(\theta) = \frac{1}{R} \sum_{r=1}^{R} P_{ij}(\beta_i^{r|\theta}),$$

where R is the number of draws and $SP_{ij}(\theta)$ is an unbiased estimator of P_{ij}. The simulated log-likelihood function is then simply $\sum_{ij} ln(SP_{ij}(\theta))$ (Revelt & Train, 1998). The estimated parameters are those that maximize this log-likelihood function. The simulated probability is an unbiased estimate of the true probability, but the log of this probability is not and introduces a bias. This bias, however, can be decreased by increasing the number of repetitions.

4.4 Deriving willingness-to-pay

The validity of a measurement of WTP often depends on the method used (Krystallis et al., 2006b). Respondents often overestimate their purchase likelihood due to costs being taken less into consideration than in real life situations. This results in an overestimation of consumer's WTP and an underestimation of price sensitivity (Krystallis et al., 2006b). By using conjoint analysis, a more realistic situation is created and the chances of overestimating WTP are decreased. As previously stated, in this simple conjoint analysis, the two product attributes that differ per product are price (in euros) and ecolabel.

In Section 4.3, it was already briefly discussed that there are two ways to calculate WTP. The first is the conventional method, which is called the preference space model, where WTP for an attribute can be derived by assuming a distribution for the coefficients and taking the ratio of the attribute coefficient to the monetary coefficient. WTP can thus be derived by taking the ratio of two randomly distributed terms. The methodology on deriving WTP follows from Train & Weeks (2005). Consider the utility function as previously introduced in equation (10),

$$U_{ijt} = -\alpha_j p_{ijt} + \beta_j x_{ijt} + \varepsilon_{ijt}. \tag{13}$$

Here, the error term is assumed to be an iid (independent and identically distributed) extreme value term that consists of a scale parameter for every decision maker i. The extreme value distribution is based on limiting distributions for maximums or minimums (e.g. extreme values) of iid random variables in samples of increasing size (Kotz & Nadarajah, 2000). According to Hansen (2020), the extreme value distribution can be described as follows: if we have a statistical distribution $p(x)$ with cumulative probability

$$P(x) = \int_{-\infty}^{x} p(x')dx',$$

which corresponds to the probability of finding a number smaller or equal to x, and we draw N numbers from this distribution and only record the largest number, and then repeat this M times, what would be the distribution of these M largest numbers in the limit when $M \to \infty$?

The scale of utility is determined by the variance of the error term, which is given by σ_j. This is an individual-scale parameter, meaning that it is different for every decision maker due to factors that are known to the decision maker but unknown to the researcher (Train & Weeks, 2005). For the standard logit, the variance of the distribution of the error term in the utility function is fixed and equal to $\pi^2/6$, which is the variance researchers traditionally use (Train, 2009). Train & Weeks (2005) state that if the error term in the utility function in equation (13) represents factors known to the decision maker but unknown to the researcher, there may be little need for the scale parameter to vary across decision makers. He therefore suggests fixing the scale parameter such that the variance of the error term is equal to the variance from the standard logit. This allows for normalizing the scale of utility, which is equivalent to normalizing the variance of the error terms. Fixing the variance is done by dividing the utility function in equation (13) by the variance of the error term σ_j, which results in the standard logit error term with fixed variance $\pi^2/6$:

$$U_{ijt} = -\lambda_j p_{ijt} + \theta_j x_{ijt} + \eta_j z_i + \varepsilon_{ijt}.$$

Here, $\lambda_j = \alpha_j/\sigma_j$ and $\theta_j = \beta_j/\sigma_j$ etc. Now, the different coefficients will be correlated unless the scale parameter does not vary over individuals (Hole & Kolstad, 2012).

Given equation (13), WTP can be derived as follows: $w_j = \beta_j/\alpha_j$. Intuitively, this can be explained by thinking of the concept of WTP. WTP presents how much someone is willing to pay for a unit upgrade in a certain product attribute. In other words: what is the amount of money that has the same value in utility as the product attribute? To obtain this, the change in utility has to be equal to 0, and given a one unit upgrade in the product attribute, the price has to increase by the ratio of the utility values of the attribute over the price.

According to Train & Weeks (2005), the conventional formulation of utility could, however, result in distributions of WTP that are heavily skewed. He therefore suggests to use the definition of WTP to rewrite the utility equation. Given $w_j = \beta_j/\alpha_j$, the utility function can be rewritten as follows:

$$U_{ijt} = -\alpha_j p_{ijt} + (\alpha_j w_j)' x_{ijt} + \varepsilon_{ijt}. \tag{14}$$

Equation (14) is called the utility model in WTP space. Now, the coefficients of the model already represent the desired WTP measure. The variation in WTP is now distinguished from the variation in the price coefficient and independent of scale. While this approach is said to produce more realistic WTP results, the most suitable approach for calculating WTP may differ for different types of data.

As mentioned before, the `logitr()` package in R allows for estimating both a mixed logit based on utility in the preference space, as well as a mixed logit based on utility in the WTP space, where the estimation in preference space refers to the standard approach of deriving WTP from the ratio of two coefficients (the ratio of the attribute coefficient to the price coefficient).

5 Results

5.1 Analysing differences in purchase likelihood

In the first part of the survey, respondents were shown the organic or non-organic type of a food product at random, and asked to indicate how likely they would be to purchase this product. The data from this part of the survey allow us to determine whether the purchase likelihoods between the organic and non-organic variant of the different food products are significantly different. This is done with the help of Mann-Whitney U-tests. This method is preferred over the T-test since the assumption of normality is not found to hold for the data. A significant difference between the purchase likelihood for organic versus non-organic food products can be seen as an indication that further research into the drivers of organic food preferences is useful. In Table 1, the results of the Mann-Whitney U tests for all five food products are found. Here, the null hypothesis states that there is no difference between the purchase likelihood of organic products and non-organic products.

Table 1 shows that for spaghetti and milk, a significant difference in purchase likelihoods between organic and non-organic products is found. For all other products, the null hypothesis that there is no difference in purchase likelihood for the two types of products cannot be rejected. This implies that the effect of ecolabels differs for different types of food products and that there should be no one-size-fits-all policy for organic food products.

Table 1: Mann-Whitney U test results for all five food products

Product	P-value	Reject null hypothesis
Chicken	0.057	No
Spaghetti	0.001	Yes
Milk	0.001	Yes
Tomatoes	0.296	No
Banana's	0.202	No

5.2 Ordered logit results

An ordered logit is performed to see how different consumer characteristics affect the probability of a consumer falling under a certain consumer type. Here, the consumer types as explained in Section 4.1 relate to the respondent's behaviour in relation to the current prices and price changes of the organic and non-organic variant of a food product. The consumer types range from 1 to 4 and are in natural order, with consumer type 1 never preferring the organic alternative, consumer type 2 only preferring the organic alternative when its price is 10% lower than the supermarket price, consumer type 3 only preferring the organic alternative when equal to the supermarket price or lower, and consumer type 4 always preferring the organic alternative.

An ordered logit is performed for all five different food products (chicken, milk, tomatoes, bananas, and spaghetti), which allows for identifying important differences in preferences between food products based on consumer characteristics. For every logit, the predictor variables are *Gender, Age, Education, Employment, Household size*, and *Income*. Here, *Gender* takes on the value 1 if the respondent is male and 2 if the respondent is female. The variable *Age* is divided into nine age scales as displayed in Section 3, where 1 corresponds to the lowest age scale (<12 years) and 9 to the highest age scale (>75 years). *Education* can take on 9 values as displayed in Section 3, where 1 corresponds to the lowest educational level (no schooling) and 9 to the highest level (PhD or Doctorate Degree). For all questions, answers corresponding to "Prefer not to say" are removed from the data. The variable *Employment* is divided into 7 categories: (1) unable to work (2) unemployed and not looking for work (3) unemployed and looking for work (4) retired (5) student (6) self-employed and (7) employed for wages. The variable *Household size* is divided into 7 categories, and the value 1 corresponds to the smallest household size (1 person household) and 7 to the largest household size (more than six person). Finally, the variable *Income* is divided into 13 scales with 1 being the lowest income scale (between €0 – €9999) and 13 the highest income scale (€500,000 or more). The data is divided into a train and test set where the train set consists of 75% of the total number of observations, and the test set of the other 25%. The test set is used to evaluate the performance of the model on new data.

Table 2 shows the results of the ordered logit for all five food products. The first column corresponds to chicken, the second to spaghetti, the third to milk, the fourth to tomatoes and the fifth to bananas. It can be seen that most demographic variables are statistically insignificant at all significance levels, meaning they do not have enough explanatory power to explain changes in the dependent variable (consumer type). Age is found to be statistically significant at all significant levels for chicken, spaghetti and tomatoes, and significant at the 0.05 level for bananas. The employment status of the respondent is found to have a significant effect on consumer type for tomatoes at the 0.05 level. Furthermore, income is found to significantly affect consumer type for spaghetti, milk and tomatoes at the 0.1 level and the 0.01 level for tomatoes.

The coefficients are displayed in log odds. By taking the exponent of the coefficients, the odds ratios can be obtained. When taking the exponent of the coefficient of the variable *Age* for chicken, an odds ratio of approximately 1.40 is obtained. This tells us that for every increase in age category, the odds of preferring the organic food alternative, even with a higher price, are multiplied by 1.40.

Table 2: Ordered logit results for (1) chicken (2) spaghetti (3) milk (4) tomatoes (5) bananas

	Dependent variable:				
	Consumer type				
	(1)	(2)	(3)	(4)	(5)
Gender	0.374	0.093	0.725**	0.186	0.119
	(0.382)	(0.383)	(0.363)	(0.479)	(0.364)
Age	0.336***	0.311***	0.028	0.311***	0.211**
	(0.104)	(0.105)	(0.095)	(0.114)	(0.100)
Education	0.120	0.089	0.067	−0.051	0.124
	(0.101)	(0.106)	(0.093)	(0.122)	(0.101)
Employment	0.093	−0.065	0.077	0.401**	0.041
	(0.162)	(0.155)	(0.155)	(0.200)	(0.151)
Household size	0.094	0.043	0.170	0.021	0.094
	(0.157)	(0.157)	(0.129)	(0.183)	(0.137)
Income	0.040	0.113*	0.093*	0.232***	0.069
	(0.067)	(0.063)	(0.057)	(0.080)	(0.062)
Observations	151	138	152	139	144
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 2 shows that being female has a positive significant effect on consumer type for milk, and age has a positive significant effect on the dependent variable for chicken, spaghetti, tomatoes and bananas. A positive effect means that an increase in this variable results in a higher probability of the consumer falling in a more "organic-oriented" consumer type. Education is not found to have a statistical significant effect on consumer type for all five food products. Employment status is only found to have a positive significant effect on consumer type for tomatoes. Household size is also found to be statistically insignificant for all five food products. Furthermore, income is also found to have a significant positive effect on consumer type for most food products, which is in line with the discussed studies in Section 2 that found a positive effect of disposable income on organic consumption behaviour.

In a second analysis, all variables without any significant effect for all five food products are excluded. The results of this analysis can be found in Table 3. It can be seen that removing the insignificant variables has very little effect on the previously found results and the model can therefore be regarded as relatively robust.

Table 3: Ordered logit results for (1) Chicken (2) Spaghetti (3) Milk (4) Tomatoes (5) Bananas

	Dependent variable:				
	Consumer type				
	(1)	(2)	(3)	(4)	(5)
Gender	0.294	0.051	0.632*	0.193	0.020
	(0.374)	(0.379)	(0.355)	(0.472)	(0.353)
Age	0.299***	0.289***	−0.016	0.309***	0.170*
	(0.094)	(0.097)	(0.088)	(0.109)	(0.092)
Employment	0.104	−0.060	0.071	0.384*	0.059
	(0.161)	(0.155)	(0.151)	(0.197)	(0.149)
Income	0.059	0.124**	0.118**	0.230***	0.085
	(0.062)	(0.058)	(0.053)	(0.074)	(0.059)
Observations	151	138	152	139	144
Note:			*p<0.1; **p<0.05; ***p<0.01		

The limitations of the data that were previously discussed become evident when using the ordered logit to predict consumer types. For all five food products, the models classify all consumers as either type 1 (never preferring the organic product) or type 4 (always preferring the organic product). This shows that it may be more useful to focus on identifying and predicting strictly organic vs. non-organic consumers, as will be done in the next section with the help of random forests and support vector machine classification models.

5.3 Predicting consumer choices

The first research question as introduced in Section 1 focuses on developing a model that predicts whether a consumer will choose the organic or non-organic alternative of a food product against the current supermarket prices. This is done with the help of random forests and support vector machine (SVM) classification models. For every method, several models were tested and compared against each other. These models and their results are discussed below.

5.3.1 Random Forests

In this section, random forests are used to classify consumers into organic and non-organic consumers. Because of the relatively small number of observations in the data, leave-one-out cross-validation is used to improve the estimated performance of the model without introducing a bias in the model. By repeatedly leaving one observation out of the training data and validating on this observation, a more accurate estimate of the true underlying performance of the model can be obtained.

Two different models are developed for all five food products, resulting in a total of ten classification models. The first model is referred to as the simplified model and includes only demographic variables as predictor variables. The second model is called the extensive model and includes all demographic variables, purchase likelihood values for all food products except the food product for which the prediction is made, and the respondent's views on their behaviour towards organic food products as predictor variables. The latter corresponds to (1) the respondent's awareness of ecolabels on food products, (2) the respondent's view on how the presence of ecolabels affects their purchasing behaviour and (3) the respondent's main reason for purchasing organic food products.

The data is again divided into a train and test set where the train set contains 75% of the observations. The test set is used to evaluate the performance of the model. For the purpose of the analysis, all missing observations are removed from the data.

Table 4: Confusion matrices for random forest models on test set, chicken data

(a) Simplified model

		Reference		
		NO	**O**	Total
Prediction	**NO**	32	9	41
	O	4	5	9

(b) Extensive model

		Reference		
		NO	**O**	Total
Prediction	**NO**	31	8	39
	O	5	6	11

Table 4 shows the results for the two models on the test set for chicken. Here, "NO" stands for non-organic consumer, and "O" for organic consumer. The results show that the extensive model performs better on the test set, with an accuracy of approximately 74%. Due to the class imbalance in the data, it is more informative to look at the balanced accuracy. Balanced accuracy is measured by adding the sensitivity metric and the specificity metric and dividing by two. For the extensive model, the sensitivity metric, which measures the share of true positive cases detected, is equal to approximately 0.86. The specificity metric however, which reflects the rate of true negatives, is equal to approximately 0.43. This results in a balanced accuracy rate of approximately 0.65 or 65%. For the simplified model, the accuracy lies at 74% as well, but the balanced accuracy lies slightly lower at 62%. To analyze the predictive power of the indicated purchase likelihood values of the other food products, a third analysis is performed in which only the demographic variables and the respondent's views of their behaviour towards organic food products are included. This model is called the control model. Surprisingly, the control model results in an accuracy of 78% and a balanced accuracy of 72%. This indicates that purchase behaviour (approximated by indicated purchase likelihood) does not improve predictive power and may even lead to misbeliefs on the purchase behaviour of other products. However, while this may hold for chicken, this does not necessarily have to be true for other food products.

Table 5 shows the confusion matrices for the simplified and extensive model on the test set for bananas. Both the simplified and the extensive model are found to perform worse than the models for chicken. The simplified model achieves the highest accuracy, with an accuracy of 68% and a balanced accuracy of 67%. The extensive model reaches an accuracy of 66% and a balanced accuracy of 64%.

Table 5: Confusion matrices for random forest models on test set, banana data

(a) Simplified model

		Reference		
		NO	**O**	Total
Prediction	**NO**	22	10	32
	O	6	12	18

(b) Extensive model

		Reference		
		NO	**O**	Total
Prediction	**NO**	23	12	35
	O	5	10	15

The control model, which includes demographic variables and data on the views of the respondent towards organic food products, is also applied to the banana data. This model is found to perform even worse than the simplified and extensive model, with an accuracy 58% and a balanced accuracy of only 57%. For bananas, it can therefore be concluded that the demographic characteristics of the consumer alone are relatively good predictors of organic purchase behaviour.

Table 6: Confusion matrices for random forest models on test set, milk data

(a) Simplified model

		Reference		
		NO	**O**	Total
Prediction	**NO**	23	6	29
	O	8	13	21

(b) Extensive model

		Reference		
		NO	**O**	Total
Prediction	**NO**	24	10	34
	O	7	9	16

Table 6 shows the confusion matrices for the simplified and extensive model on the test set for milk. Again, the simplified model is found to perform better than the extensive model. The accuracy of the extensive model is 66%, with a balanced accuracy of 62%. The simplified model achieves an accuracy of 72% with a balanced accuracy of 71%, a much higher performance than the extensive model. This performance cannot be improved by leaving out the purchase likelihoods in the extensive model, which leads to an accuracy of 71% and a balanced accuracy of 67%. The same conclusion of bananas can therefore be drawn for milk: consumer demographics have relatively high predictive power over the purchase behaviour of consumers.

Table 7: Confusion matrices for random forest models on test set, tomato data

(a) Simplified model

		Reference		
		NO	**O**	Total
Prediction	**NO**	31	9	40
	O	3	7	10

(b) Extensive model

		Reference		
		NO	**O**	Total
Prediction	**NO**	32	10	42
	O	2	6	8

Table 7 shows the results for tomatoes. Here, the two models perform almost equally well, with both models obtaining an accuracy of 76%. However, the simplified model achieves a slightly higher balanced accuracy of 67%, while the extensive model reaches a balanced accuracy of 66%. In line with the results for chicken, the control model performs significantly better, with an accuracy of 78% and a balanced accuracy of 72%. This indicates that for this particular product as well, information on the purchase decisions for other food products does not result in a better prediction of consumption behaviour.

Table 8 shows the results of the simplified and extensive model for spaghetti. The Table shows that both do not perform very well, since a large part of the organic consumers in the test set are misclassified. While the extensive model has a slightly higher accuracy (68%) than the simplified model (66%), the extensive model does have a lower balanced accuracy (62%) than the simplified model (63%). In line with the results for chicken and tomatoes, the model that performs best for spaghetti is the control model, which results in an accuracy of 74%, and a balanced accuracy of 69%.

Table 8: Confusion matrices for random forest models on test set, spaghetti data

(a) Simplified model

		Reference		
		NO	**O**	Total
Prediction	**NO**	24	9	33
	O	8	9	17

(b) Extensive model

		Reference		
		NO	**O**	Total
Prediction	**NO**	27	11	38
	O	5	7	12

For all food products, the model that performs best is used to evaluate the importance of the different predictor variables. Table 9 shows an overview of the best performing models for all five food products and their corresponding accuracy and balanced accuracy rates. As mentioned before, the third model, which is a variation of the extensive model in which the purchase likelihoods are excluded as predictor variables, is referred to as the control model.

The best performing models and their accuracy rates are summarized in Table 9. The Table shows that for none of the food products, the extensive model is found to be the best performing model. This indicates that purchase likelihoods of other products, which can be used as a proxy for actual purchase behaviour, do not provide a good indication of the purchase behaviour of other products in relation to organic versus non-organic choices.

Table 9: Best performing random forest models for all five food products

Product	Best performing model	Accuracy (*balanced accuracy*)
Chicken	control model	78% (*72%*)
Bananas	simplified model	68% (*67%*)
Milk	simplified model	72% (*71%*)
Tomatoes	control model	78% (*72%*)
Spaghetti	control model	74% (*69%*)

5.3.1.1 Variable importance

In this section the variable importance of the different models is evaluated and the results are compared to previous findings on the importance of different consumer characteristics. All variables are assigned an impact score based on a scale from 0 to 100. The importance is based on the mean decrease in Gini Index. The more the Gini Index decreases for a feature, the more important this feature is.

Table 10 shows that for food products where the control model is found to be the best performing model, the variable “Choice influence”, relating to the extent to which the consumer is influenced by the presence of ecolabels, is often found to be a very important predictor of organic consumption behaviour. In nearly all best performing models, income and age are among the top five most important predictor variables, thus supporting hypothesis 3 introduced in Section 2. Furthermore, of all food products for which the best performing model is the control model, the consumer’s main reason for organic consumption (for example environmental concerns or animal welfare) is found to be an important determinant as well. This is in line with hypothesis 2, which states that personal reasons and motives are important determinants of organic consumption. Education is also regularly found to be an important determinant, thus supporting hypothesis 4, stating that educational level is an important determinant organic consumption. Overall, household size is not found to be an important predictor in the best performing random forest models.

Table 10: Variable importance of the best performing random forest models for all five food products

Chicken		Bananas		Milk		Tomatoes		Spaghetti	
variable	**value**	**variable**	**value**	**variable**	**value**	**variable**	**value**	**variable**	**value**
Label awareness	100.00	Age	100.00	Income	100.00	Choice influence	100.00	Choice influence	100.00
Choice influence	62.25	Income	97.43	Age	66.01	Age	76.75	Age	87.30
Income	59.45	Education	95.98	Education	38.63	Education	29.74	Income	54.51
Reason organic choice	60.00	Household size	45.02	Gender	30.35	Income	10.45	Reason organic choice	40.59
Education	20.41	Education	0.00	Household size	0.00	Household size	10.38	Label awareness	28.36

5.3.2 Support Vector Machines

In this section, consumers are classified into organic and non-organic consumers with the help of SVM models. For all five food products, the simplified, extensive and control model are applied again. However, in this case, four different types of SVM are tested for all three models: linear SVM with fixed cost or regularization parameter (held constant at 1), linear SVM with tuned regularization parameter, non-linear SVM with radial kernel and finally, non-linear SVM with polynomial kernel. For each food product, the best performing model is then again taken for further analysis and compared to the best performing random forest model. Before running the model, all variables are scaled to a [0,1] range in order to assure that the optimal hyperplane is not influenced by the scale of the input features. Again, leave-one-out cross validation is used to tune the hyperparameters of the model and to choose the optimal regularization parameter.

Table 11 shows the results for chicken. The simplified model classifies all observations as non-organic, hence the balanced accuracy of only 50% for all models. The extensive, linear SVM models perform slightly better with a balanced accuracy of 68%. In line with the random forest model, the control model performs best for chicken. Here, two out of the four models obtain an accuracy of 82% and a balanced accuracy of 72%.

Table 11: Accuracy and balanced accuracy (BA) on test set for all SVM models, chicken data

	Linear with fixed cost	Linear with tuned cost	Non-linear radial	Non-linear polynomial
Model	**Accuracy (BA)**	**Accuracy (BA)**	**Accuracy (BA)**	**Accuracy (BA)**
Simplified	72% (50%)	72% (50%)	72% (50%)	72% (50%)
Extensive	76% (68%)	76% (68%)	72% (50%)	74% (60%)
Control	72% (50%)	82% (72%)	80% (71%)	82% (72%)

Table 12 shows the results for bananas. It can be seen that in line with the random forest model, the simplified model performs best. In specific, the simplified SVM model with radial kernel achieves the highest accuracy, with an accuracy of 68% and a balanced accuracy of 69%. All other models perform significantly worse.

Table 12: Accuracy and balanced accuracy (BA) on test set for all SVM models, banana data

	Linear with fixed cost	Linear with tuned cost	Non-linear radial	Non-linear polynomial
Model	**Accuracy (BA)**	**Accuracy (BA)**	**Accuracy (BA)**	**Accuracy (BA)**
Simplified	62% (61%)	64% (63%)	68% (69%)	58% (59%)
Extensive	64% (62%)	58% (56%)	66% (63%)	58% (55%)
Control	60% (59%)	64% (63%)	60% (57%)	58% (59%)

Table 13 shows the results for the milk data. In contrast with the random forest model, the best performing model here is the extensive model corresponding to the SVM model with radial kernel. This model achieves an accuracy of 70% with a balanced accuracy of 66%. This is lower than the accuracy the optimal random forest model of milk reached. It can also be seen that all simplified models for milk classify all consumers as non-organic consumers, with balanced accuracy rates of 50%.

Table 13: Accuracy and balanced accuracy (BA) on test set for all SVM models, milk data

	Linear with fixed cost	Linear with tuned cost	Non-linear radial	Non-linear polynomial
Model	**Accuracy (BA)**	**Accuracy (BA)**	**Accuracy (BA)**	**Accuracy (BA)**
Simplified	62% (50%)	62% (50%)	62% (50%)	62% (50%)
Extensive	60% (59%)	64% (62%)	70% (66%)	64% (62%)
Control	66% (64%)	64% (62%)	64% (63%)	66% (63%)

Table 14 shows the results for the tomato data. In line with the random forest models, the control model is found to perform best based on balanced accuracy. This model is an SVM model with radial kernel. The accuracy and balanced accuracy are, however, slightly lower than the optimal random forest model for this data.

Table 14: Accuracy and balanced accuracy (BA) on test set for all SVM models, tomato data

	Linear with fixed cost	Linear with tuned cost	Non-linear radial	Non-linear polynomial
Model	**Accuracy (BA)**	**Accuracy (BA)**	**Accuracy (BA)**	**Accuracy (BA)**
Simplified	68% (50%)	68% (50%)	74% (66%)	68% (50%)
Extensive	74% (64%)	74% (64%)	78% (66%)	74% (61%)
Control	68% (50%)	72% (60%)	76% (71%)	72% (60%)

Table 15 shows the results for spaghetti and in line with the random forest model, the control model performs best for this data. Surprisingly, the linear model with the regularization parameter held constant at 1, performs better than all other types of SVM.

Table 15: Accuracy and balanced accuracy (BA) on test set for all models, spaghetti data

	Linear with fixed cost	Linear with tuned cost	Non-linear radial	Non-linear polynomial
Model	**Accuracy (BA)**	**Accuracy (BA)**	**Accuracy (BA)**	**Accuracy (BA)**
Simplified	68% (64%)	66% (61%)	62% (53%)	62% (52%)
Extensive	62% (56%)	66% (61%)	72% (66%)	72% (67%)
Control	74% (68%)	72% (65%)	72% (66%)	72% (66%)

Table 16 shows the best performing models for all five food products and their corresponding accuracy and balanced accuracy percentages. Surprisingly, milk is the only food product for which the extensive model performs best, and banana is the only food product for which the simplified model achieves the highest accuracy. All other best performing models are control models. It should also be noted that the non-linear SVM with radial kernel performs best for 3 out of the 5 models. For chicken, two SVM models achieved the same (high) performance: the linear model with tuned cost, and the non-linear model with polynomial kernel.

Table 16: Best performing SVM models for all five food products

Product	Best performing model	Type of SVM	Accuracy (BA)
Chicken	control model	linear with tuned cost, polynomial	82% (72%)
Bananas	simplified model	non-linear radial	68% (69%)
Milk	extensive model	non-linear radial	70% (66%)
Tomatoes	control model	non-linear radial	76% (71%)
Spaghetti	control model	linear with fixed cost	74% (68%)

5.3.2.1 Variable importance

In this section the variable importance of the best performing models for all five food products are again evaluated and compared to previous findings on the importance of different consumer characteristics. The variables are again assigned a relative impact score based on a scale from 0 to 100. For the chicken data, the SVM model with radial kernel is chosen for evaluating variable performance as this model was found to perform best for almost all other food products, and it can therefore be assumed that this type of SVM works well with the survey data in general.

Table 17: Variable importance values for all five food products - SVM

Chicken		Bananas		Milk		Tomatoes		Spaghetti	
variable	**value**	**variable**	**value**	**variable**	**value**	**variable**	**value**	**variable**	**value**
Label awareness	100.00	Income	100.00	Choice influence	100.00	Choice influence	100.00	Choice influence	100.00
Choice influence	92.73	Education	94.72	Label awareness	80.27	Age	66.05	Label awareness	67.79
Income	49.68	Age	65.59	Reason organic choice	47.57	Income	61.49	Income	66.30
Age	46.26	Household size	38.53	Purchase likelihood chicken	40.13	Label awareness	53.88	Age	55.00
Household size	13.08	Gender	0.00	Gender	22.3	Reason organic choice	33.09	Gender	44.57

Table 17 shows that of all demographic variables in the data, *Income* and *Age* are two variables that often recur in the top five most important predictor variables. Furthermore, *Choice influence* and *Label awareness* are two variables that when included in the model, are often the most important predictors. Here *Choice influence* relates to the question whether respondents feel that the presence of an ecolabel affects their purchase behaviour. *Label awareness* relates to the question whether respondents are aware of the presence of ecolabels on food products. The results are similar to the random forest models and coincide with the hypotheses formulated in Section 2. A more extensive comparison between the optimal random forest and SVM models is performed in the next section.

5.3.2.2 Model comparisons: random forests vs. support vector machines

In this section, the two best performing models of both machine learning methods are compared to each other based on accuracy, balanced accuracy and variable importance. Table 18 shows the two best performing models for all five food products. It can be seen that for three of the five food products, random forests perform better than SVM. This is especially the case for milk, where the difference in balanced accuracy between the two methods is the greatest. For all five food products, the best performing models achieve a balanced accuracy of around 70%.

Table 18 also shows that for all food products except milk, the best performing model is the same for random forests as for SVM. For these models, we can therefore further compare the best performing models of the two methods based on variable importance.

Table 18: Best performing models: random forests vs. SVM

	Random forests		Support vector machines	
Product	**Best performing model**	**Accuracy (BA)**	**Best performing model**	**Accuracy (BA)**
Chicken	control model	78% (72%)	control model	82% (72%)
Bananas	simplified model	68% (67%)	simplified model	68% (69%)
Milk	simplified model	72% (71%)	extensive model	70% (66%)
Tomatoes	control model	78% (72%)	control model	76% (71%)
Spaghetti	control model	74% (69%)	control model	74% (68%)

Tables 10 and 17 show that for chicken, the three most important predictor variables for both methods are *Label awareness*, *Choice influence* and *Income*. *Age* and *Household size* are the fourth and fifth most important predictors for the SVM model, whereas the variable indicating the main reason for choosing organic and *Education* are the fourth and fifth most important variables in the random forest model.

For bananas, *Age*, *Income*, *Education* and *Household size* are in the top four most important variables for both methods. Only the fifth most important variable differs, but these variables are both assigned a score of 0, thus being of relatively little importance to the prediction performance of the models. For tomatoes, more differences can be found. For both methods, the variables *Choice influence* and *Age* are the most important predictor variables. For the SVM model, income is a relatively important variable in the prediction with a score of 61.49, while for the random forest model, this variable is assigned a score of only 10.45 and *Education* is a more important predictor variable. For the SVM model, *Label awareness* and *Reason organic choice* are relatively important predictor variables but these variables are not part of the five most important variables for the random forest model.

The fifth food product is spaghetti. For both machine learning methods, *Choice influence* is the most important variable for the prediction. *Label awareness*, *Age* and *Income* also appear in the top five most important variables for both methods. *Gender* is found to be a relatively important predictor in the SVM model but not in the random forest model. In the latter, the respondent's indicated main reason for choosing organic is found to be a more important predictor.

5.4 Mixed logit results

In this part of the analysis, a mixed logit is performed to see how the price levels and attributes of a certain food product affect the consumer's choice for the organic and non-organic alternative of this product. Furthermore, the mixed logit allows us to analyze the WTP for a product with ecolabel. As discussed in Section 4.4, there are two different methods for deriving WTP. The first is the conventional method, which is called WTP under the preference space model. In this model, WTP is simply derived by taking the ratio of the attribute coefficient to the monetary coefficient. The second model is called the WTP space model. Here, the coefficients of the model already represent the WTP measure. In this study, both the preference space and the WTP space model are used and the results are compared to each other.

As mentioned in Section 3.1, the survey data collected for the purpose of this research showed some important limitations that affect the informativeness of the results of the mixed logit. Therefore, only the data of bananas is used, as this data was shown to have the most variation in consumer types. In a second analysis, data of the Liss Panel Data Archive is used to support the previous findings. Here, only the choice questions between organic and non-organic variants are included in the analysis to allow for better comparisons. In this research, all parameters are calculated under the normal distribution, assuming that the price coefficient can take on both positive and negative values.

Table 19 shows the results for the mixed logit in the preference space under the normal distribution for bananas and the Liss Panel chicken breast data. The Table shows that the coefficients of both price and ecolabel are insignificant for both data sets. This indicates that these attributes do not have enough explanatory power to explain the choice for organic food products among consumers. Despite the insignificant coefficients, WTP for the products under the preference space model is calculated. This is done by taking the ratio of the ecolabel coefficient to the price coefficient. Table 20 reports the mean and standard deviation of the distribution of WTP for the two models. The WTP for organic bananas and organic chicken breast are respectively €3.20 and €0.80 according to this model. However, the Table also shows us that the standard deviations of both these parameters are extremely high and that the results shown in the Table are therefore extremely unreliable and not very informative.

Table 19: Mixed logit results under normal distribution for bananas (survey data) and chicken breast (Liss Panel data)

	Bananas			Chicken breast (Liss Panel data)		
	Mean	**St. dev**	**P-value**	**Mean**	**St. dev**	**P-value**
Price	-0.194	1.820	0.915	-0.237	0.484	0.625
Ecolabel	0.626	7.762	0.936	0.184	0.891	0.836

Table 20: WTP for bananas and chicken breast under preference space model.

	Bananas			Chicken breast (Liss Panel data)		
	Mean	**St. dev**	**P-value**	**Mean**	**St. dev**	**P-value**
Lambda	0.194	1.815	0.915	0.235	0.482	0.623
Ecolabel	3.231	50.702	0.949	0.778	226.055	0.997

It may therefore be more informative to look at the results of the WTP space models in Table 21. Here, the coefficients already represent the WTP measures. For bananas, the results are very different from the previous model and the WTP estimate lies much lower with a much smaller standard deviation, therefore indicating to be more reliable and informative. For chicken breast, WTP under this model is similar to the WTP reported under the preference space model. Furthermore, all estimates are now found to be statistically significant. It can therefore be assumed that on average, consumers are willing to pay only 9 cents extra for bananas with an ecolabel, and only 68 cents for ecolabeled chicken. These estimates are relatively low, especially since the actual price differences between ecolabeled and non-ecolabeled products are much bigger. This could possibly be explained by the limitations of the data, with the majority never preferring the organic product due to the higher price tag. Both models, do however, show that the WTP for organic chicken is higher than the WTP for organic bananas. This could possibly be explained by the fact that meat products have a much bigger impact on the environment than fruits and vegetables.

Table 21: Results of WTP space model for bananas and chicken breast

	Bananas			Chicken breast (Liss Panel data)		
	Mean	**St. dev**	**P-value**	**Mean**	**St. dev**	**P-value**
Lambda	0.014	0.000	0.000	0.018	0.000	0.000
Ecolabel	0.093	0.012	0.000	0.678	0.001	0.000

As previously discussed, when only focusing on those choice questions involving organic versus non-organic products for the Liss Panel data, 70% of the choice questions relate to respondents choosing the non-organic variant over the organic variant. However, when taking a closer look, this percentage changes significantly based on the price difference between the different variants of chicken. The data showed that when the price of organic chicken jumps from €3.73 to €5.72, the share of respondents preferring non-organic chicken increases from 65% to 80%. Given these findings, an additional mixed logit is performed which focuses solely on choice questions involving organic variants of chicken with prices up to €5.72. This could yield more informative results, since the data showed that most preference changes occur in the €2.71 - €5.72 price range for organic chicken. Focusing on these choice questions did, however, not result in a statistically significant coefficient of the distribution of WTP in the preference space model. Furthermore, the results under the WTP space are found to remain more or less the same. This indicates that despite the larger number of observations and variation in the data, the price and ecolabel attributes of the products are not found to have a statistically significant effect on the choice for organic products under the preference space model.

6 Conclusions and discussion

6.1 Main findings

This study focuses on revealing consumer preferences for organic food products with the help of survey data. The study does so by focusing on two separate research questions:

1. *Is it possible to predict organic consumption behaviour, and what are the main determinants of organic consumption behaviour?*
2. *What is the effect of organic food labels on willingness-to-pay for different types of food products?*

In order to obtain an idea of the expected results, Mann-Whitney U tests are performed for five different food products (chicken, spaghetti, milk, tomatoes and bananas), to determine whether there is a significant difference in average purchase likelihood (based on a 5-point Likert scale) for organic and non-organic products. The products that are found to have significant differences in purchase likelihood are milk and spaghetti. These products are also the two products with the smallest price differences between their organic and non-organic variant. However, these prices were not revealed yet in the part of the survey this analysis is based on. These results already indicate that there is no one-size-fits-all policy for organic food products.

After performing the Mann-Whitney U tests, ordered logits for all five food products were performed to analyze the effect of consumer demographics on consumer type, where 1 corresponded to the least organic-oriented consumer (never buying the organic food product), and 4 corresponded to the most organic-oriented consumer (always buying the organic product, even with a 10% increase in price). The results show a positive and significant effect of age and income on consumer type for almost all five food products. However, because of limitations of the data, the model classifies all respondents as either consumer type 1 or 4. These results indicated that focusing on predicting organic and non-organic consumers would be more likely to yield insightful results.

Predicting consumer types is done with the help of random forest and support vector machine models. Consumers are classified as organic or non-organic consumers based on their behaviour when facing the choice between the organic and non-organic alternative of a food product against the current supermarket prices. The results show that respondent demographics and personal views and norms on organic consumption behaviour are the most important indicators of consumer type. This is in line with hypotheses 2 and 3 introduced in Section 2. Furthermore, the purchase likelihoods of other food products were not found to carry predictive power in predicting organic consumption behaviour. The random forest model was found to perform better for 3 of the 5 food products, based on accuracy and balanced accuracy. The average accuracy of all best performing models lies around 74% and the average balanced accuracy, which is more insightful given the imbalance in the data, lies around 71%. These results serve as important indications of the value of consumer data in predicting purchase behaviour.

In the final part of the analysis, a mixed logit is performed to measure WTP for food products with an ecolabel. This is done for bananas (from the survey data) and chicken (from the Liss Panel data) under both the preference space and the WTP space model. The conventional preference space model was not found to yield statistically significant results. Furthermore, when computing WTP under the preference space model, the standard deviations are extremely high and the results are therefore not very robust. The WTP model on the other hand, was found to yield significant results. The model showed an average WTP for ecolabeled bananas of only 9 eurocents. For the Liss Panel data, a WTP of approximately €0.70 for organic chicken breast relative to regular chicken breast was found. This is very low considering the price points of ecolabeled and regular chicken, but could possibly be explained by the fact that the great majority of the respondents always prefers the non-organic alternative.

This study reveals that consumers are overall, not very prone to price changes in organic food alternatives. However, the study also shows that the majority of consumers can be classified as either fully "organic" or "non-organic", and that these consumer classes can be predicted relatively accurately. Furthermore, a relatively large part of the consumers can be classified as "organic" (around 40% on average across food products). These results have very important implications for the bio-food sector. Being able to predict organic consumers allows for far more effective targeting strategies which can save companies enormous amounts of marketing costs. Furthermore, if the share of "organic" consumers keeps growing, switching from conventional to organic farming can be very profitable in the long-run, especially in combination with more effective marketing efforts from both supermarkets and food processor companies.

6.2 Discussion and limitations

The main limitations of this study involve the data and lack of variation in the data, as was illustrated in Section 3. In the simplified conjoint analysis, the great majority of the sample either always preferred the organic product, or never preferred the organic product. This indicates that the great majority of consumers is not prone to price changes in organic food products and has a strong preference for one of the two. Because of this, measurements of WTP were unreliable and not very informative, especially under the conventional preference space model. Another limitation of the data is that in the conjoint analysis, the price points of the non-organic alternatives remained the same for all 3 choice questions. This was done intentionally in order to limit the number of choice questions and avoid survey fatigue. However, if more choice questions would have been included and the price points of the non-organic alternative would have varied as well, this would have perhaps led to more informative results and implications. This remains questionable however, given that the Liss Panel data did involve variations in price points of non-organic variants, and the results of the preference space model were not found to be more informative. Nevertheless, consumption switching behaviour based on price changes in both organic and non-organic products should be researched further in future research.

A second limitation of the survey data is that the respondent's answers do not necessarily have to mimic actual purchase behaviour. There is a possibility that part of the respondents who were found to be "organic-oriented", would not opt for the organic option in a real-life situation because the product would for example, not be the first product they would see in the supermarket.

A third limitation of the survey data is the elicitation issue that could have arisen due to the survey design. In the first part of the survey, consumers are asked to indicate the purchase likelihood for a food product that could either have or not have an ecolabel. Respondents who notice that certain products have ecolabels and other products do not, may have anticipated the purpose of the study, which could have steered their answers towards favoritism of organic products in the second part of the survey. However, as it was found that most respondents do not prefer the organic alternative in this part of the survey, it is very unlikely that this is an issue. Nevertheless, it is important to note and could still have played a role in this study.

It should also be noted that the data from the Liss Panel Archive used in this study were collected in 2012. As mentioned before, recent studies have shown that due to the COVID-19 pandemic, consumer preferences have changed drastically and consumers now attach more and more value to food safety and a healthy lifestyle overall. It is therefore likely that the results from this data are not externally valid. This is important to take into account, as it is very likely that the actual WTP for organic chicken lies significantly higher when using new and recent data.

While the data are not suitable for predicting different consumer types based on their switching behaviour when faced with price changes, it is suitable and quite successful in predicting organic and non-organic consumers. Furthermore, the models allow for evaluating feature importance in the prediction. For future research, these models could be extended by applying model agnostic methods to study the direction of the effects of the most important variables on organic consumption choices.

References

Aschemann-Witzel, J., & Zielke, S. (2017). Can't buy me green? a review of consumer perceptions of and behavior toward the price of organic food. *Journal of Consumer Affairs*, *51*(1), 211–251.

Biau, G. (2012, April). Analysis of a random forests model. *J. Mach. Learn. Res.*, *13*(null), 1063–1095.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. Retrieved from `http://dx.doi.org/10.1023/A%3A1010933404324` doi: 10.1023/A:1010933404324

Bruin, J. (2021). *Ordinal logistic regression — r data analysis examples.* Retrieved 2021-05-10, from `https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/`

Bunte, F. H., van Galen, M. A., Kuiper, W. E., & Tacken, G. (2010). Limits to growth in organic sales. *De Economist*, *158*(4), 387–410.

Ecovia Intelligence. (2020). *Organic foods getting coronavirus boost.* Retrieved from `https://www.ecoviaint.com/organic-foods-getting-coronavirus-boost/,urldate={2021-04-17}`

Foster, C., Green, K., Bleda, M., et al. (2007). Environmental impacts of food production and consumption: final report to the department for environment food and rural affairs.

Galarraga Gallastegui, I. (2002). The use of eco-labels: a review of the literature. *European Environment*, *12*(6), 316-331. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/eet.304` doi: https://doi.org/10.1002/eet.304

Genuer, R., Poggi, J.-M., & Tuleau, C. (2008). *Random forests: some methodological insights.*

Gil, J. M., Gracia, A., & Sanchez, M. (2000). Market segmentation and willingness to pay for organic products in spain. *The International Food and Agribusiness Management Review*, *3*(2), 207–226.

Gunn, S. (1998). *Support vector machines for classification and regression* (Project Report). Retrieved from `https://eprints.soton.ac.uk/256459/` (Address: Southampton, U.K.)

Hansen, A. (2020). The three extreme value distributions: An introductory review. *Frontiers in Physics*, *8*, 533.

Hemmelskamp, J., & Brockmann, K. L. (1997). Environmental labels—the german 'blue angel'. *Futures*, *29*(1), 67-76. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0016328796000663` doi: https://doi.org/10.1016/S0016-3287(96)00066-3

Hole, A., & Kolstad, J. (2012). Mixed logit estimation of willingness to pay distributions: a comparison of models in preference and wtp space using data from a health-related choice experiment. *Empirical Economics*, *42*(2), 445-469. Retrieved from `https://EconPapers.repec.org/RePEc:spr:empeco:v:42:y:2012:i:2:p:445-469`

Jha, V., Vossenaar, R., Zarrilli, S., on Environmental Policies, W., & Access, M. (1993). Ecolabelling and international trade..

Klein, B., & Leffler, K. B. (1981). The role of market forces in assuring contractual performance. *Journal of political Economy*, *89*(4), 615–641.

Kotz, S., & Nadarajah, S. (2000). *Extreme value distributions: theory and applications.* World Scientific.

Krystallis, A., & Chryssohoidis, G. (2005). Consumers' willingness to pay for organic food: Factors that affect it and variation per organic product type. *British food journal*.

Krystallis, A., Fotopoulos, C., & Zotos, Y. (2006a). Organic consumers' profile and their willingness to pay (wtp) for selected organic food products in greece. *Journal of international consumer marketing*, *19*(1), 81–106.

Krystallis, A., Fotopoulos, C., & Zotos, Y. (2006b). Organic consumers' profile and their willingness to pay (wtp) for selected organic food products in greece. *Journal of International Consumer Marketing*, *19*(1), 81-106. Retrieved from `https://doi.org/10.1300/J046v19n01_05` doi: 10.1300/J046v19n01_05

Laroche, M., Bergeron, J., & Barbaro-Forleo, G. (2001). Targeting consumers who are willing to pay more for environmentally friendly products. *Journal of consumer marketing*.

Martínez, R. E. B., Ramírez, N. C., Mesa, H. G. A., Suárez, I. R., Trejo, M. d. C. G., León, P. P., & Morales, S. L. B. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista médica de la Universidad Veracruzana*, *9*(2), 19–24.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *42*(2), 109–142. Retrieved from `http://www.jstor.org/stable/2984952`

McFadden, D., & Train, K. (2000). Mixed mnl models for discrete response. *Journal of Applied Econometrics*, *15*(5), 447-470. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/1099-1255%28200009/10%2915%3A5%3C447%3A%3AAID-JAE570%3E3.0.CO%3B2-1` doi: https://doi.org/10.1002/1099-1255(200009/10)15:5⟨447::AID-JAE570⟩3.0.CO;2-1

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, *10*(1), 1–16.

Ministerie van Landbouw, Natuur en Voedselkwaliteit. (2019). *Xiv landbouw, natuur en voedselkwaliteit en f diergezondheidsfonds rijksbegroting 2020.* Retrieved 2021-04-21, from `https://www.rijksoverheid.nl/documenten/begrotingen/2019/09/17/xiv-landbouw-natuur-en-voedselkwaliteit-rijksbegroting-2020`

Mons, G. (2020). *Omschakelen naar bio: verkijk je niet op de kosten.* Retrieved 2021-04-21, from `https://www.akkerwijzer.nl/artikel/233811-omschakelen-naar-bio-verkijk-je-niet-op-de-kosten/#:%7E:text=Ruwweg%20kost%20omschakelen%205.000%20euro,areaal%20circa%20500%20hectare%20uitbreidt`

Padel, S., & Foster, C. (2005). Exploring the gap between attitudes and behaviour: Understanding why consumers buy or do not buy organic food. *British food journal*.

Paul, J., & Rana, J. (2012). Consumer behavior and purchase intention for organic food. *Journal of consumer Marketing*.

Peattie, K. (1995). *Environmental marketing management: Meeting the green challenge.* Financial Times Management.

Revelt, D., & Train, K. (1998, nov). Mixed logit with repeated choices: Households' choices of appliance efficiency level. *Review of Economics and Statistics*, *80*(4), 647–657. Retrieved from https://doi.org/10.1162%2F003465398557735 doi: 10.1162/003465398557735

Rijksdienst voor Ondernemend Nederland. (2020). *Vergroeningsbetaling 2020.* Retrieved 2021-05-21, from https://www.rvo.nl/subsidie-en-financieringswijzer/vergroeningsbetaling

Rijksoverheid. (2021a). *Biologische landbouw: arealen en veestapels, 2011-2020.* Retrieved from https://www.clo.nl/indicatoren/nl0011-biologische-landbouw#:%7E:text=In%202018%20bedraagt%20het%20aandeel,landbouwsector%20in%20ons%20land%20klein

Rijksoverheid. (2021b). *Landbouwsubsidies.* Retrieved 2021-04-22, from https://www.rijksoverheid.nl/onderwerpen/landbouw-en-tuinbouw/landbouwsubsidies

Roheim, C. A., Asche, F., & Santos, J. I. (2011). The elusive price premium for ecolabelled products: Evidence from seafood in the uk market. *Journal of Agricultural Economics*, *62*(3), 655-668. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1477-9552.2011.00299.x doi: https://doi.org/10.1111/j.1477-9552.2011.00299.x

Shafie, F. A., & Rennie, D. (2012). Consumer perceptions towards organic food. *Procedia - Social and Behavioral Sciences*, *49*, 360-367. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877042812031308 (Proceedings of the 1st National Conference on Environment-Behaviour Studies, 1nCEBS, FAPS, UiTM, Shah Alam, Malaysia, 14–15 November, 2009) doi: https://doi.org/10.1016/j.sbspro.2012.07.034

Skal. (2021). *Biologische sector groeit 2,5% in 2020.* Retrieved 2021-04-29, from https://www.skal.nl/nieuws/biologische-sector-groeit-2-5-in-2020

Smith, N. C. (2014). *Morality and the market (routledge revivals): Consumer pressure for corporate accountability.* Routledge.

The Comprehensive R Archive Network. (2021). *The random parameters (or mixed) logit model.* Retrieved 2021-04-21, from https://cran.r-project.org/web/packages/mlogit/vignettes/c5.mxl.html

Train, K. (2009). *Discrete choice methods with simulation.* Cambridge University Press. Retrieved from https://EconPapers.repec.org/RePEc:cup:cbooks:9780521766555

Train, K., & Weeks, M. (2005). Discrete choice models in preference space and willingness-to-pay space. In R. Scarpa & A. Alberini (Eds.), *Applications of simulation methods in environmental and resource economics* (pp. 1–16). Dordrecht: Springer Netherlands. Retrieved from https://doi.org/10.1007/1-4020-3684-1_1 doi: 10.1007/1-4020-3684-1_1

Van Ravenswaay, E., & Hoehn, J. P. (1991). *Consumer willingness to pay for reducing pesticide residues in food: Results of a nationwide survey.* Department of Agricultural Economics, Michigan State University.

7 Appendix

7.1 Appendix A: Survey questions

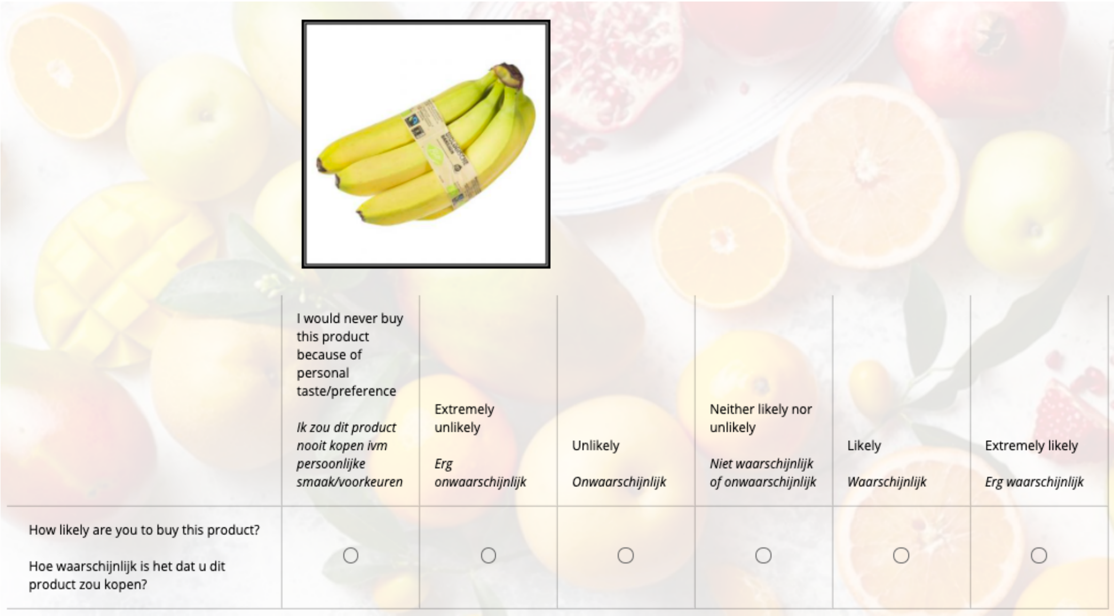
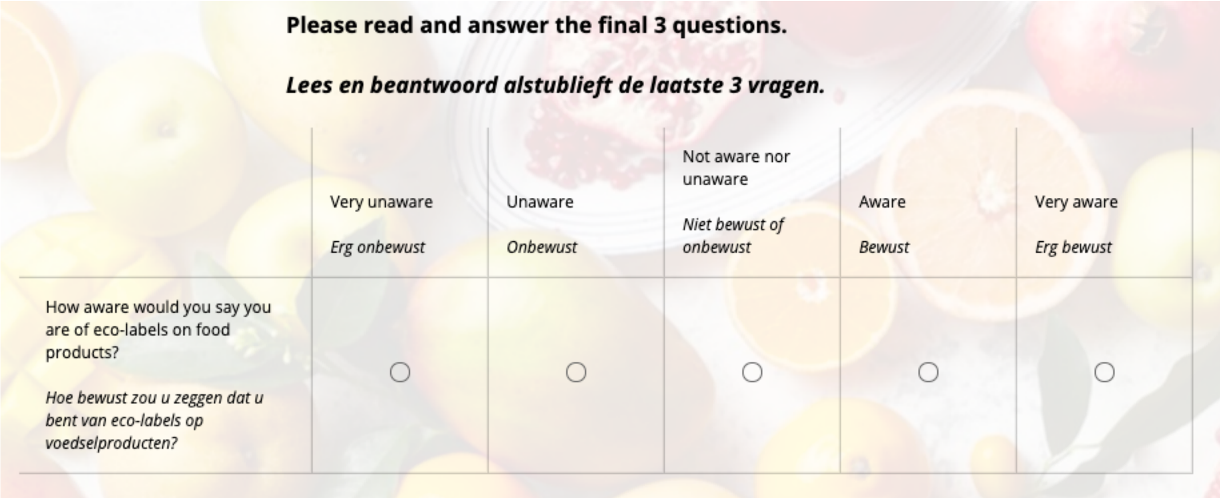


Figure 9: Example question purchase likelihood

Figure 10: Example question conjoint analysis



Please read and answer the final 3 questions.

Lees en beantwoord alstublieft de laatste 3 vragen.

	Very unaware *Erg onbewust*	Unaware *Onbewust*	Not aware nor unaware *Niet bewust of onbewust*	Aware *Bewust*	Very aware *Erg bewust*
How aware would you say you are of eco-labels on food products? *Hoe bewust zou u zeggen dat u bent van eco-labels op voedselproducten?*	○	○	○	○	○

Figure 11: Survey question about ecolabel awareness

Please state to what extent you agree with the following statement.

Geef alstublieft aan in hoeverre u het eens bent met de volgende stelling.

	Completely disagree *Helemaal niet eens*	Disagree *Oneens*	Neither agree nor disagree *Niet eens of oneens*	Agree *Eens*	Completely agree *Helemaal mee eens*
The presence of eco-labels on food-products affects my choice to buy a product. *De aanwezigheid van eco-labels op een voedselproduct beïnvloedt mijn keuze om een product te kopen*	○	○	○	○	○

Figure 12: Survey question about whether the presence of ecolabels affects the consumer's choice to buy a certain product

What would be the main reason for preferring a biological/eco-friendly product over a non-biological product?

Wat zou de grootste reden zijn voor het kiezen van een biologisch/milieuvriendelijk product boven een niet biologisch product?

To be more environmentally friendly
Om meer milieuvriendelijk te zijn

To be more animal friendly
Om meer diervriendelijk te zijn

To be more healthy
Om gezonder te zijn

Because biological products are more socially acceptable
Omdat biologische producten meer maatschappelijk geaccepteerd zijn

Because biological products are tastier
Omdat biologische producten lekkerder zijn

I would never prefer a biological product over a non-biological product
Ik zou nooit een biologisch product boven een niet-biologisch product kiezen

Other, please specify:
Anders, graag specificeren:

Figure 13: Survey question about the respondent's main reason for choosing an organic food product

For the link to the survey, please click **here**.