ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Master Thesis Data Science and Marketing Analytics

The Sustainability Attitude-Behaviour Gap

A Machine Learning Analysis of a Preferendum among Youth

Name student: David Brummer
Student ID number: 541909

Supervisor: Prof. dr. M.G. de Jong
Second assessor: Prof. dr. P.J.F. Groenen
Date final version: 23rd of July 2020

The Sustainability Attitude-Behaviour Gap

A Machine Learning Analysis of a Preferendum among Youth

David Brummer

Abstract

Through the comparison of different methods, this research aims to uncover which factors seem to influence the attitude-behaviour gap in a young target audience, namely Gen Z. After reducing dimensions using PCA, this paper compares a white-box method, the elastic net logistic regression to a black-box method, the random forest algorithm. Both models include all PCA-obtained components in which altruism, family wealth, social norms, social desirability bias and trust in various sources are captured. Besides controlling for these factors, the models simultaneously control for various background factors and externally merged variables of which age, education, water risk score and SDG index show to be important predictors in both obtained models. Due to the preference of predicting the underrepresented class correctly, the established model using a penalised logistic regression is favoured. Additionally, the increased level of interpretation of the coefficients of this white-box method adds to the arguments to favour this method above the random forest algorithm, which predicts better overall. The findings of the analysis have implications for policymakers, educational institutions and other organisations that battle climate change through policy-interventions that include, but are not limited to artificially inducing norms, creating awareness of consequences through the water footprint concept and highlighting the parental role model status.

Keywords: attitude-behaviour gap, sustainable water consumption, Gen Z, principal components analysis, regularization, logistic regression, random forest

Contents

Introduction

"You have stolen my dreams and my childhood with your empty words. And yet I'm one of the lucky ones. People are suffering. People are dying. Entire ecosystems are collapsing. We are in the beginning of a mass extinction, and all you can talk about is money and fairy tales of eternal economic growth. How dare you!", Greta Thunberg (2019) addresses her speech toward international leaders at the U.N. Summit for Climate Action. Although the science has been around for more than three decades, countries have looked and are still looking away. While the current plans to cut emissions by 50% already seem unattainable for many nations, future generations are relied upon to solve the multi-dimensional climate problem by coming up with innovative technologies. The speech continues that to have a reasonable chance of keeping global temperatures rising below 1.5 °C, all countries around the globe need to stay below to their nation's emission budget. This 16-year old climate activist's message is an appeal to unite nations in the fight against climate change. Her approach is to wake world leaders up, confronting them with the consequences of their lack of action, representing future generations that will be left with the consequences of the status quo.

NASA (Shaftel, 2019) defines climate change as follows: "... the long-term change in the average weather patterns that have come to define Earth's local, regional and global climates". The Earth has been through many naturally occurring cycles of climate change. The main actor in this process is carbon dioxide (CO_2), one of the most well-known greenhouse gases (GHGs). GHGs have the property of absorbing heat, known as infrared radiation, emitted from the Earth's surface. The warming of the earth increases in the magnitude of the concentrations of these gases in the atmosphere (Nordhaus, 2019). However, the observed size of temperature change is extremely likely (more than 95% probability) driven by unsustainable human behaviour (Shaftel, 2019). More specifically, the GHG concentrations in the atmosphere are especially driven by burning carbon-based fossil fuels on a global scale (Nordhaus, 2019). The observed growth of the atmospheric concentrations comes down to 40% over the last 200 years, with more than 20% growth in the past 50 years, since the Industrial Revolution, warming the Earth by approximately 1 °C. To put this into context, the average temperature rise since the last ice age was 1 °C, which occurred over a period of around 7.000 years, which is 35 times slower than the changes observed over the last 200 years (The Royal Society, 2017). Temperature increases have widespread consequences, materialised in the form of oceans warming up, melting of Arctic sea ice, which in turn together result in the rise of the sea level and the increase in the frequency of extreme weather situations (The Royal Society, 2020). Additionally, scientists have shown their concern about potential 'tipping points'. This occurs when the damage has become irreversible, with the melting ice sheets and large changes in the ocean circulation being considered important global tipping points (Nordhaus, 2019). The consequences of climate change will drive the evolution of social systems, with all its social, economic

and political implications, while affecting the well-being of individuals in society (Rosa and Dietz, 2012). Moreover, Nordhaus (2019) argues that potential impacts are considered to be the largest for those natural systems with high vulnerability to climate-sensitive systems, such as rain-fed agriculture, natural ecosystems and forest fires and erosion.

The severity of climate change and the role of mankind with respect to this multidimensional problem is recognised by a growing collection of nations, world leaders, policymakers and scientists. The IPCC (Intergovernmental Panel on Climate Change) was founded to regularly assess climate change scientifically and create an overview of (future) impacts and provide input for mitigation policies. IPCC illustrates in their Fifth Assessment Report that the economic sectors Energy Supply, Agriculture, Forestry & Other Land Use (AFOLU), Industry, Transport, Other Energy and Buildings contribute to global greenhouse gas emissions, in descending order of responsibility (IPCC, 2014). At the end of 2015, the United Nations Framework Convention on Climate Change (UNFCCC) established a landmark agreement in Paris during the COP21, an annually held Conference of Parties (COP). This agreement aims to battle climate change by increasing efforts and investments worldwide for a sustainable future. A year later the Paris Agreement was signed, consisting of more than 150 parties' nationally determined contributions (NDCs) (UNFCCC, 2018; Berrang-Ford et al., 2019). However, to achieve the clearly articulated goals of adaptation and reducing (vulnerability to) climate change, many have criticised the Agreement of missing a uniform approach to assessing progress in established adaptation commitments. Without mechanisms or frameworks in place to uniformly assess both governments' accountability and adaptation, there is no clear way to decide on the impact of The Agreement, limiting its efficacy (Nordhaus, 2019; Berrang-Ford et al., 2019). Complementarily, Nordhaus (2019) argues that such a policy can only be effective under the following conditions: the gravity of the impacts of global warming on both mankind and the ecosystem needs to be understood, nations should unanimously establish a policy on carbon prices, action should be coordinated globally and innovative technologies are necessary to battle this by mankind driven threat to the Earth.

Thunberg (2019) confronts the world leaders with the rapidly depleting carbon budget, confirming the findings above and implying the urgency of lasting radical change to how this threat is treated. The problem of GHG emissions, however, is two-sided. Through the last decades population, wealth, consumption, technology, institutional agreements and culture have been considered candidate drivers of climate change. Consumption is considered the main driver of atmospheric GHG emissions, which is heavily interrelated with (the technology used in) manufacturing, transport and the disposal of generated waste. Moreover, this consumption is then again influenced by the growing global population and the level of affluence (Rosa and Dietz, 2012). Individual households, therefore, take part in the responsibility of reducing GHG emissions by showing pro-environmental

behaviours, such as waste management, minimising energy consumption and consuming water sustainably.

Many researchers have tried explaining behaviour using different frameworks (Ajzen, 1991; Schwarz, 1977; Stern, Dietz, Abel, Guagnano and Kalof, 1999) and including many different factors ranging from background factors such as demographics, education, personality traits and values (Gifford and Nilsson, 2014), to emotions (Onwezen, Antonides and Bartels, 2013), future-perspectives (Rabinovich, Morton and Postmes, 2010) and goals (Steg, Bolderdijk, Keizer and Perlaviciute, 2014). However, an interesting question is raised, namely: does a pro-environmental attitude always result in pro-environmental behaviour? Besides behaviour, this closely-related phenomenon has been well-documented, defined as the attitude-behaviour gap (Rajecki, 1982; Kollmuss and Agyeman, 2002; de Leeuw, Valois, Ajzen and Schmidt, 2015).

Due to the ever-growing presence of climate change, the managerial relevance of uncovering potential drivers of mentioned gap lies in the field of policymakers and educational institutions, helping create more effective policies and education to close the gap and move to a more sustainable world. The youth are the changemakers of the future, highlighting the need for research to provide input for policies by examining the process of sustainable behaviour formation (Grønhøj and Thøgersen, 2012). Additionally, this research contributes to the existing literature in the following ways: this paper focuses on one specific sustainable behaviour, namely sustainable water consumption. Adding to that, the analysis focuses on a very young target audience, namely Gen Z. Finally, the analysis focuses on young people ($N = 3.440$) from eight different countries, ranging from India to The Netherlands. The following research question is, therefore, examined:

- *Which factors drive the attitude-behaviour gap with respect to sustainable water consumption in young adults?*

Firstly, literature is reviewed, where two additional questions are answered, namely:

- *Which psychological factors determine one's attitude and behaviour toward sustainability?;*
- *Which demographic differences are found with respect to sustainable behaviour?*

Secondly, data processing is described. Moving on, the methods used for empirical research are explained, after which the empirical results are presented. Finally, this paper ends with the conclusion, where the findings are evaluated and limitations and ideas for future research are discussed.

Literature

Climate change and water scarcity

Since the Industrial Revolution in the 1970s, climate change has progressively gotten worse, with the Earth warming faster than in the 200 years before (The Royal Society, 2017). Consequently, since the 1980s research with respect to water scarcity has attracted much attention both politically and publicly (Liu et al., 2017). The Earth's warming has large effects on the climate and has many consequences on the macro, meso and micro level; The World Economic Forum (2015) mention water-related crises as the single largest risk for mankind considering its plausible effects on a global level. Consequently, Jaeger *et al.*(2017) confirm that climate change will have many consequences for water shortages worldwide and will force the human population to adjust; due to water's vital role in many natural and human systems, societies worldwide will be afflicted. Facts are that at this moment approximately 4.0 billion people, approximately 65% of Earth's inhabitants, are yearly confronted with acute water scarcity for more than a month (Mekonnen and Hoekstra, 2016), and on average 2.35 billion people live under these severe circumstances for at least four months a year, worldwide, which are only projected to increase substantially in the next 30 years (Gosling and Arnell, 2013; Liu et al., 2017; Naumann et al., 2018).

The growing global population in combination with economic development have progressively put more pressure on Earth's resources, due to its growing demand (Liu et al., 2017). Consequently, increasing water scarcity due to population growth, changing climate and growing incomes, has large effects on human-systems, affecting food production, economic development and result in perishing ecosystems (Watkins, 2006). Water scarcity and in turn its consequences, therefore, form the most pressing challenges for a sustainable future of humanity and the preservation of entire ecosystems (Hoekstra and Wiedmann, 2014; Howard, Calow, Macdonald and Bartram, 2016; Rosa, Chiarelli, Rulli, Dell'Angelo and D'Odorico, 2020). Naumann *et al.*(2018) find that continuing at the present rate of global warming, exacerbated by mankind, deficits in supply and demand will increase by a magnitude of five and those areas most at risk of droughts will be affected by once-in-a-century droughts, every two to five years.

Governance

No wonder that the UN incorporated water scarcity as an important point in their SDGs which were adopted in 2015, aiming to reduce the magnitude of the population affected by water scarcity (UN, 2015). On a country level, this translates into experiencing what is projected to become an issue for an increasing amount of countries worldwide. Four years ago in South Africa, Cape Town, the city was becoming aware of water

scarcity, resulting from years of drought. In the next two years, the city faced this increasing problem which would result into Day Zero, meaning that the dam levels of the city would come to a worrying point of 13.5%, forcing the city to impose a policy of rationing unto its inhabitants, materialising the risks of climate change for countries worldwide (Millington and Scheba, 2020). Its long-term strategy would focus on (encouraging) restriction of water consumption, implementation of taxation and augmenting its current water systems (Millington and Scheba, 2020). Other strategies found in big U.S. cities include encouraging adoption of durable alternatives, restrictions and education programmes (Kang, Grable, Hustvedt and Ahn, 2017). Capping water consumption, increasing efficiency of a wide variety of water-use, and augmenting distribution of water resources forms a crucial role in the policies of those countries confronted with water scarcity (Mekonnen and Hoekstra, 2016; Duran-Encalada, Paucar-Caceres, Bandala and Wright, 2017).

Consumer responsibility

The current state of the world is a human-modified one, called the Anthropocene, in which society actively shapes the availability of water, where demand depends on amount of people, living standards, and how efficient water is used (Van Loon et al., 2016). Since the 1900s, water consumption has increased by a magnitude of four, increasing the population affected by water scarcity from 0.24 billion people to 3.8 billion in the 2000s (Kummu et al., 2016). This implies that, besides top-down approaches, consumers play a significant role in the overall water demand (Klöckner, 2013; Hoekstra, Chapagain and Zhang, 2015); Craig, Feng and Gilbertz (2019) highlight this necessity of consumer empowerment to influence efforts to mitigate the risks and adapt to the challenges resulting from global warming. Complementarily, consumers hold the key to change, since they potentially form the largest driver towards adaptation, due to their leverage over businesses, investors and governments (Haida, Chapagain, Rauch, Riede and Schneider, 2019).

Theory of Planned Behaviour

To understand the consumer in this multidimensional playing field, the driving forces behind its consumption patterns need to be examined; more specifically, to fully understand and effectively change behaviour one has to uncover the underlying factors that are determining the behaviour (Klöckner, 2013). Widespread research has been done in the fields of psychology and sociology to uncover those factors influencing and, therefore, determining one's behaviour. One of the most prominent cognitive modelling theories is The Theory of Planned Behaviour (TPB; Ajzen, 1991). Ajzen (1991) hypothesizes that behaviour flows from one's intention to behave in a certain way, and in turn, that intention is determined by three factors, namely perceived behavioural control, subjective norm and finally, attitude. Klöckner (2013) explains

these determinants as the degree to which people have the opportunity and ability to perform a certain behaviour, the product of perceived social pressure and the willingness to comply, and the attitude towards that behaviour, respectively. Generally speaking, one's intention to behave in a certain way is positively related to the attitude, subjective norm toward that behaviour, and magnitude of perceived behavioural control toward the behaviour. Complementarily, it is expected that different situations affect the relative significance of each of the determinants with regards to intention (Ajzen, 1991). Ajzen shows that each of the determinants, in turn, is formed by beliefs and that their salience affects the importance of the resulting factors; he assumes that beliefs about the behaviour influence one's attitude, normative beliefs influence one's subjective norms and control beliefs form the underlying basis for one's perceived behavioural control. Finally, Ajzen concludes that additionally controlling for moral obligation besides the three determinants could add additional explanatory power to the model (Ajzen, 1991).

Norm-Activation Model and Value-Belief-Norm Theory

Another approach for looking at sustainable behaviour is through the Norm-Activation Model (NAM; Schwartz, 1977), which has been shown to explain altruistic behaviour. This theory assumes there to be four determinants that activate one's moral obligation, or activated personal norm before one behaves in line with it. The four determinants are the following: awareness of need and consequences, acknowledgement of one's responsibility and finally the belief that the person is capable of behaving in a certain way (Schwartz, 1977; Klöckner, 2013), which is similar to Ajzen's (1991) perceived behavioural control in TPB.

Fast forward to the late 1990s, Stern *et al.* (1999) build upon the NAM and propose their Value-Belief-Norm Theory (VBN). Comparing it to the NAM, where there are four conditions in parallel, the VBN considers the conditions to activate each other; before someone can be responsible, the person in question needs to know the consequences of certain behaviour. In the pro-environmental context, Stern *et al.*(1999) consider the New Ecological Paradigm to be a measure for mankind's acknowledgement of their influence on the natural world, recognising that mankind is bringing it out of balance. This paradigm again is influenced by one's values on altruism, traditionality, egoism and openness to change. This chain reaction, from personal values to NEP to awareness of consequences, acknowledgement of responsibility leads to the activation of one's personal norms which in turn translate into (pro-environmental) behaviour (Stern et al., 1999).

Additional factors

A large body of research has been done on the empirical validity of the models and many researchers find the different theories to be valuable. Klöckner (2013) sees the theories as complementary in the sense that VBN and NAM cannot fully explain behaviour just through one's moral obligation (activated personal norm) and need other mediating variables to explain behaviour, but on the other hand, this is exactly what TPB is missing. Additionally, what all three theories miss is the degree of habituation of old behaviour that could interfere with performing pro-environmental behaviour (Knussen and Yule, 2008; Klöckner, 2013; Carrington, Neville and Whitwell, 2014; Ajzen and Dasgupta, 2015).

Retrospectively speaking Ajzen (2011) reflects on criticism received on TPB throughout the years: many have argued that TPB is too rational, to which Ajzen replies that the beliefs that create the attitude, subjective norm and perceived behavioural control are not formed in an unbiased fashion, implying these might be based on irrational foundations. Moreover, Ajzen poses that there are other determinants influencing beliefs, that determine either of the factors that make up to be one's intention. There are many other potential background factors, spanning from values, personality traits, demographics or exposure to external sources of information, implying controlling for these additional variables could prove fruitful (Ajzen and Fishbein, 2005; Ajzen, 2011; de Leeuw et al., 2015); Gifford and Nilsson (2014) have found that a large range of personal and social factors or combinations of these factors influence sustainable behaviour, including experiences from childhood, education, one's character, sense of control, values, views on politics and the world, goals they might have, degree of taking responsibility, biases, place attachment, age, gender, religion, cultural and ethnic variations, urban versus rural differences, norms, social class and spatial proximity to an affected environment.

Integrating TPB with NAT can prove to be a successful strategy in explaining pro-environmental behaviour, as Onwezen *et al.* (2013) suggest, who research the influence of self-conscious emotions pride and guilt. Onwezen *et al.* (2013) suggest that incorporating these emotions is useful, due to the fact that pride or guilt could be evoked after evaluating one's (non-)sustainable behaviour, and conclude that these emotions indeed have a mediating effect down the line; this opens up possibilities for future research to investigate the (indirect) influence of a larger range of self-conscious emotions with respect to sustainable behaviour.

Kang *et al.*(2017) confirm that beliefs are important influencers in TPB's proposed determinants; their research shows that water resource concern, utilitarian, and ecological water beliefs affect at least one of TPB's factors, which in turn influences behaviour. This research, therefore, integrates Stern *et al.*'s (1999) assumption of beliefs underlying certain behaviours. Farrow *et al.*(2017) zoom in on beliefs on a more

collective level, namely social norms, and conclude that these prove to be a valuable tool in promoting environmentally-friendly behaviour (Kinzig et al., 2013), with descriptive norms being the most consistent, explained by their salience due to being more present than injunctive norms (Cialdini, Reno and Kallgren, 1990), which has also been confirmed in the context of the parent-child relationship (Grønhøj & Thøgersen, 2012) and for high-school students (de Leeuw et al., 2015).

From salience of social norms to salient consequences: Gómez-Llanos, Durán-Barroso and Robina-Ramírez (2020) confirm Schwartz' assumption that one needs to portray awareness of consequences, before one can take responsibility by showing that the Water Footprint (WF) helps to change people's behaviour through awareness of their consumption practices, integrating both direct and indirect water use. Complementarily, the Construal Level theory of Trope and Liberman (2010) confirms that one's ability to construe becomes increasingly difficult in the psychological distance, resulting in one's interpretation becoming more abstract. There are four levels that increase psychological distance, namely temporal, spatial, social and hypothetical; this theory is in line with one's awareness of need and consequences in the sense of spatial and temporal dimensions.

Other research suggests that the integration of social identities might prove worthwhile in explaining different behaviours and that policies can leverage this knowledge by enhancing identification with certain consumer groups that behave in a certain way (Gatersleben, Murtagh and Abrahamse, 2012; Bartels and Reinders, 2016), which suggests that someone that identifies as being pro-environmental, will behave in line with this identity.

Furthermore, as one might expect time perspective influences a person behaving in line with its attitude in terms of these future-oriented behaviours (Rabinovich et al., 2010), which implies that someone that is able to take on a distant-future perspective, more likely will behave in line with its attitude towards pro-environmental behaviour, since the consequences of climate change or even its own behaviour do not lie in the short term (Trope and Liberman, 2010; Gifford and Nilsson, 2014).

Other researchers like Steg, Bolderdijk, Keizer and Perlaviciute (2014) propose a slightly different framework which finds the foundation of behaviour to flow from one's goals, hedonic, gain and normative goals. These goals are in turn affected by one's values and situational cues, which (de)activate some of these values, and influence subsequent behaviour.

Attitude-Behaviour Gap

One's intention to behave sustainably does not always translate into actual sustainable behaviour, defined as the attitude-behaviour gap, which is becoming a larger field of interest in research and has been well-documented (Rajecki, 1982; Kollmuss and Agyeman, 2002; de Leeuw et al., 2015). Rajecki (1982) proposes four barriers that influence the gap, which is in line with many of the theories discussed above: firstly, Rajecki names direct versus indirect experience, with indirect experience having a lower influence on one's behaviour, which is in line with The Construal Level Theory (Trope and Liberman, 2010). Secondly, in line with TPB (Ajzen, 1991) normative influences are argued to affect either attitude or action, which can be opposed, widening the gap. Furthermore, temporal discrepancy is argued to affect the stability of one's attitude through time; complementarily, Rabinovich *et al.*(2010) confirm that an "individual's time-perspective might also influence attitude-behaviour consistency through increasing accessibility of attitudes at the point of decision-making and the stability of attitudes over time". Finally, the gap is concluded to occur in a situation where measurement of either attitude and behaviour, results in inconsistencies due to different scopes or definitions.

Kollmuss and Agyeman (2002) conclude that there are many potential barriers explaining the gap between attitude and action, included are contradicting values and knowledge, emotional blocking of new knowledge or portraying a certain attitude, absence of internal and external incentives, environmental consciousness, insufficient feedback on behaviour, and old habits or behavioural patterns. More recently, a new framework has been proposed, called the SHIFT framework, which aims to close the gap between sustainable attitudes and behaviours, by focusing on strategies in the context of "Social influence, Habit formation, Individual self, Feelings and cognition and Tangibility" (White, Habib and Hardisty, 2019). Carrington *et al.* (2014) have researched factors influencing this gap, concluding that degree of prioritisation of ethical concerns, establishing plans or habits, one's degree of commitment and willingness to sacrifice, and shopping behaviour modes play a significant role in this gap between attitude and behaviour. Conclusively, future research should go into other potential obstacles explaining this gap (Carrington et al., 2014).

Data

The data used for this research comes from a preferendum. Respondents have several options to respond to a question. The survey is divided into two parts of which one is compulsory and one is optional. The global State of Youth preferendum has been carried out among youth and young adults. State of Youth is an organisation founded by KidsRights. The organisation focuses on increasing awareness in a young community between 13 and 24 years old. By uniting these changemakers of the world, the platform tries to enable the next generation to find their voice, and potentially create 'real change and impact, driven by youth, across the world' (State of Youth, 2019). The survey is distributed using the Facebook platform, restricted to those with English as their default language on Facebook. The research survey arose through a collaboration between KidsRights and Facebook, where the former is an academic partner of the Erasmus University Rotterdam. The preferendum has been developed by Erasmus' academics and facilitated through the Qualtrics platform. This is the first global State of Youth preferendum monitoring the voices of young people. The activation rate of the preferendum is considered to be low, with 10.900 youngsters having responded to the survey in over 60 countries worldwide. Additionally, the survey has not been completed by all 10.900 respondents.

The questions posed are both about factual details, opinions and behaviour. Firstly, the respondents are asked different demographic questions on age, living country and education. Moving on, the survey poses questions that capture the respondents' opinions on climate change, whether additional action is needed to tackle it and which of five different behaviours occurred in the respective respondents' households in the last year. Additionally, the respondents were asked how much they have learned from and trust different sources on the topic of climate change. The last questions of the mandatory part of the survey encapsulate the respondents' skillfulness in mathematics, English and whether they understood all questions posed in the survey thus far. The mandatory part of the survey ends here and asks the youngsters whether they would like to continue to answer some additional questions. In the second part the youngsters are asked how willing they are to spend money to reduce climate change, even when others do not, from which the respondents' altruistic characters can be derived. Furthermore, the next question asks the adolescents what percentage of young people in their respective countries think climate change is a serious problem and what percentage would be willing to spend money to tackle the problem, which can be viewed as a proxy for social norms. Moreover, the youngsters are asked to rate their relative knowledge compared to their peers in their respective countries, and how many extreme weather days were experienced in the last year, and how much they expect to experience in the following year. Second to last, the youngsters are asked questions on their future orientation, risk tolerance and altruism. Finally, the last question asks how privileged the respondent is in terms of wealth, whereafter the last subquestions focus on character traits, which proxy for social desirability bias.

The resulting dataset derived from the survey consists of 70 variables, 17 of which are unimportant details recorded by Facebook about when the survey was filled in, how long it took, IP address, longitude, latitude and so on. The dataset consists of 10.900 observations, but as mentioned before, not all respondents have completed the survey. Before this paper can dive into the empirical part of its research, appropriate data processing steps need to be taken.

Processing

Firstly, the 17 unnecessary variables, mentioned above, have been deleted, namely *StartDate*, *EndDate*, *Status*, *IPAddress*, *Progress*, *Duration_in_seconds*, *Finished*, *RecordedDate*, *ResponseId*, *RecipientLastName*, *RecipientFirstName*, *RecipientEmail*, *ExternalReference*, *LocationLongitude*, *LocationLatitude*, *Distribution-Channel* and *UserLanguage*. Additionally, the questions about whether they were born in the country they live in and whether they have other innovative ideas on how to tackle climate change, have been removed, as these are considered unimportant for further analysis. Moreover, since the attitude-behaviour gap of sustainable water consumption behaviour is examined, the other four sustainable behaviours, reduced meat consumption, plastic recycling, protesting against climate change and volunteering at an organisation focusing on tackling climate change are left out for further research. Additionally, all NA values in the reduced water consumption behaviour have been turned into zeros, showing that the respondents did not show sustainable water behaviour in their households, compared to observations that did show the behaviour in the past 12 months. All survey questions have been renamed with intuitive names to increase clarity and establish greater overview. Furthermore, all questions where the respondents are asked to rate their answers on a scale have been processed into numerical variables, to avoid obtaining too many factor variables and increasing interpretation of the further analysis. The questions about age, highest education, extreme weather days experienced and expected are processed into ordered factor variables, where the final questions about character traits are just factor variables. Additionally, invalid answers have been changed into the most probable answer based on the consistency of scaling of prior answers. To overcome affecting the results of the analysis, the dataset is subsetted to only include those respondents that stated to have understood all questions, whereafter the variable was excluded for further analysis. Finally, all numerical variables have been scaled, so that coefficients resulting from models are comparable, and do not affect variable importance.

Attitude-Behaviour Gap

After having processed the dataset, the data frame is subsetted to only include observations with a complete range of answers, resulting in a data frame with 3.777 observations. Since this thesis focuses on modelling

factors influencing the attitude-behaviour gap, this dependent variable has to be established. There are two questions from which the attitude can be derived. For someone to have a pro-environmental attitude, the answers to the questions about whether climate change is a serious problem and whether extra action should be taken in order to tackle the problem combined, show one's pro-environmental attitude. Two new numerical variables are created that return a one when the given answer is above four, representing a pro-environmental attitude, and a zero otherwise. Multiplying these two, into an aggregated *Attitude* variable, will return either a pro-environmental attitude (= 1), when the respondent shows to have this attitude for both of the original two variables, and will not show a pro-environmental attitude otherwise. Finally, to model the gap, by substracting the reduced water consumption behaviour from this aggregated attitude, will return the existence of a gap, represented by a one, zero otherwise. After the gap is modelled, the four variables, three on attitude and one on reduced water consumption behaviour are not included in the analysis to overcome issues that would arise through multicollinearity. 43 observations show a negative attitude-behaviour gap, where there is at least one non-pro-environmental attitude, whilst showing sustainable water consumption behaviour. These observations are excluded from the analysis since this thesis focuses on the incongruence between portraying a pro-environmental attitude, but not behaving in line with the attitude. Behaving sustainably, but not showing pro-environmental attitude is not considered a gap, at least not one that is in the scope of this research. The resulting complete data frame consists of 3.734 observations.

External data

To find out which factors influence one's attitude-behaviour gap regarding sustainable water consumption, external data could prove to be a valuable addition to the analysis. When considering variables, one must evaluate whether controlling for such a variable, would show an effect on this gap. Since the living country is available for all observations, the most logical step towards merging external data would be by doing this on the country level. However, before doing so, the data frame is subsetted to only include observations that live in a country, that has a minimum amount of observations, namely 50, resulting in a data frame of 3.440 observations. For the analysis it is only useful to consider observations of different countries, where there are a sufficient amount of observations per country, to uncover country-specific trends. Therefore, the analysis includes only nine countries, of the more than 50 originally present in the dataset, namely Brazil, India, Indonesia, Mexico, Netherlands, Nigeria, South Africa, United Kingdom and the United States. Various sources of country-specific data over 2019 have been examined and evaluated, from which two interesting variables are considered to show value for further analysis. Firstly, one can expect to show more sustainable consumption behaviour, when the thing to be consumed is more scarce. Therefore, an overall index factor

on water-related risk is added to the data frame. Hofste *et. al* (2019, Aqueduct 3.0) created a water risk framework, that aggregates 13 indicators of water risk into one overall risk indicator, for the nine countries focused on in this research. Secondly, to control for a country's stance on sustainability, which is expected to be correlated to the country's policies, media coverage, and overall compliance with the Sustainable Development Goals (SDGs), the SDG index (2019, Sustainability Development Report) per country has been merged into the data frame. Finally, after having merged all external data into the data frame, the variable living country is removed; country-specific data is more valuable than controlling for the exact country because it holds more (valuable) information. The final conceptual model that includes the variables used to model the relationship with the dependent variable can be found in the figure below.

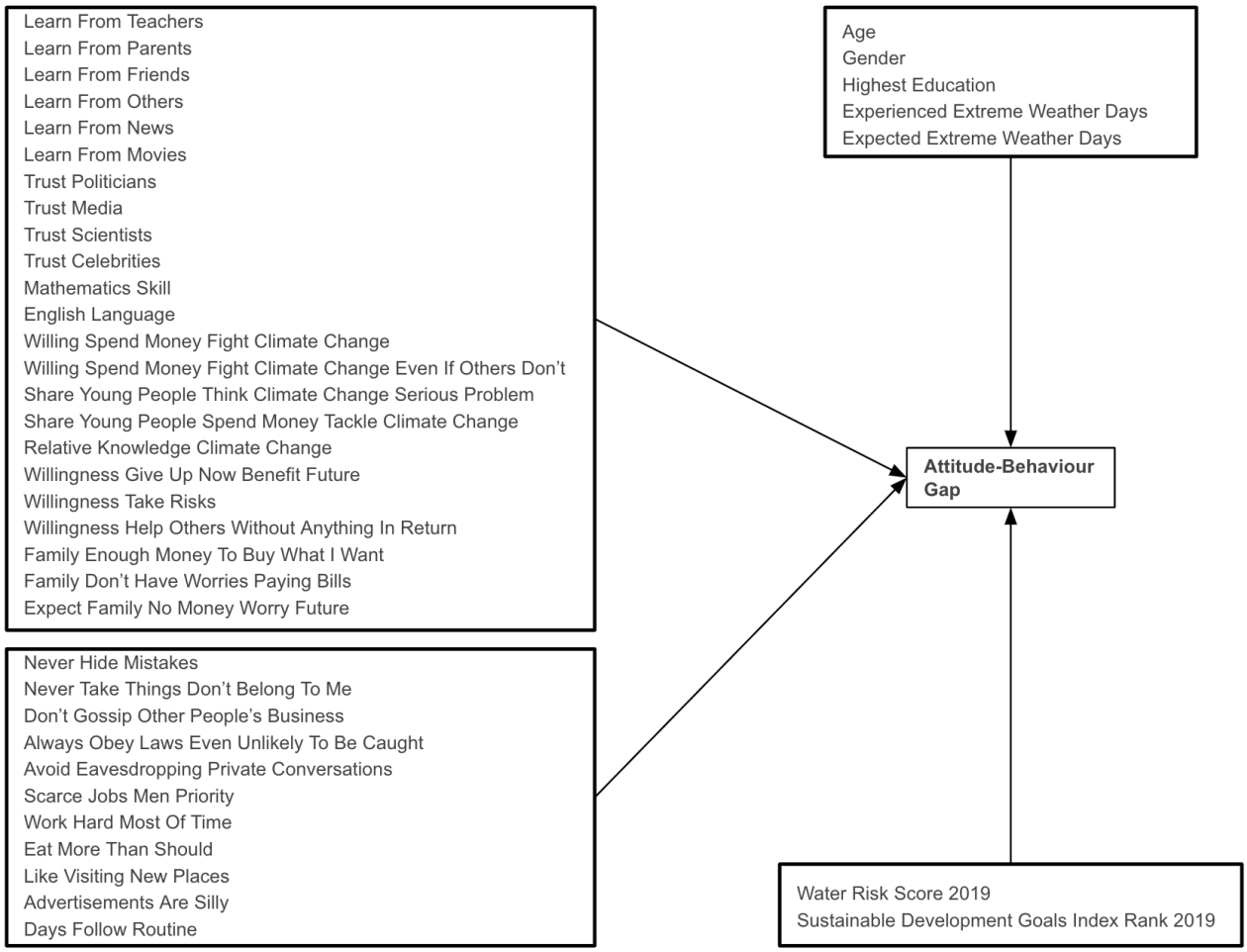

Figure 1: Conceptual model with predictor variables

Summary statistics

Numeric variables

After having processed the data for further analysis, one can examine the data. The dataset consists of 3.440 observations over 42, of which 31 numerical and 11 categorical, variables, including the variables obtained through merging external data, as mentioned above, and the attitude-behaviour gap, derived from three variables, originally present in the dataset. Firstly, one can find a table with the summary statistics of the numerical variables in Table 1 below. As one can see all numerical variables are those variables on which the respondents were asked to rate their stance or opinion, except for the final two variables. These variables, obtained through external data sources, are *WaterRiskScore2019* and *SDGIndexRank2019* and are country-specific.

Table 1: Descriptive summary statistics numeric variables

	Mean	St.Dev	Min.	Max.	Range	Skew	Kurtosis
LearnFromTeachers	4.38	1.93	1.00	7.00	6.00	-0.31	-1.18
LearnFromParents	3.97	1.99	1.00	7.00	6.00	-0.07	-1.29
LearnFromFriends	4.52	1.85	1.00	7.00	6.00	-0.48	-0.93
LearnFromOthers	4.91	1.80	1.00	7.00	6.00	-0.73	-0.49
LearnFromNews	5.73	1.54	1.00	7.00	6.00	-1.39	1.23
LearnFromMovies	4.54	1.98	1.00	7.00	6.00	-0.47	-1.05
TrustPoliticians	2.29	1.40	1.00	7.00	6.00	1.27	1.28
TrustMedia	3.43	1.69	1.00	7.00	6.00	0.25	-0.96
TrustScientists	6.01	1.16	1.00	7.00	6.00	-1.63	3.32
TrustCelebrities	3.42	1.63	1.00	7.00	6.00	0.26	-0.80
MathematicsSkill	6.59	2.39	0.00	10.00	10.00	-0.79	0.11
EnglishLanguage	8.80	1.37	0.00	10.00	10.00	-1.56	3.64
WillingSpendMoneyFightCC	5.05	1.55	1.00	7.00	6.00	-0.77	-0.01
WillingSpendMoneyFightCCEvenIfOthersDont	5.03	1.60	1.00	7.00	6.00	-0.73	-0.15
ShareYoungPplThinkCCSeriousProblem	54.51	22.29	0.00	100.00	100.00	-0.12	-0.93
ShareYoungPplSpendMoneyTackleCC	34.97	20.98	0.00	100.00	100.00	0.56	-0.29
RelativeKnowledgeCC	5.52	1.10	1.00	7.00	6.00	-0.83	0.87
WillingnessGiveUpNowBenefitFuture	8.19	1.68	0.00	10.00	10.00	-1.03	1.51
WillingnessToTakeRisks	7.85	1.78	0.00	10.00	10.00	-0.91	1.11
WillingnessHelpOthersWithoutAnythingInReturn	8.71	1.56	0.00	10.00	10.00	-1.60	3.64
FamilyEnoughMoneyToBuyWhatIWant	4.11	1.80	1.00	7.00	6.00	-0.03	-1.13
FamilyDontHaveWorriesPayingBills	3.89	1.84	1.00	7.00	6.00	0.12	-1.13
ExpectFamilyNoMoneyWorryFuture	3.86	1.85	1.00	7.00	6.00	0.08	-1.11
NeverHideMistakes	4.85	1.68	1.00	7.00	6.00	-0.43	-0.89
NeverTakeThingsDontBelongToMe	5.81	1.66	1.00	7.00	6.00	-1.43	1.05
DontGossipOtherPeoplesBusiness	5.08	1.78	1.00	7.00	6.00	-0.62	-0.78
AlwaysObeyLawsEvenUnlikelyToBeCaught	5.44	1.56	1.00	7.00	6.00	-0.91	-0.02
AvoidEavesDroppingPrivateConversations	5.18	1.81	1.00	7.00	6.00	-0.70	-0.73
ScarceJobsMenPriority	2.24	1.85	1.00	7.00	6.00	1.32	0.43
WaterRiskScore2019	2.85	1.27	0.43	4.12	3.69	-0.23	-1.65
SDGIndexRank2019	65.71	8.18	46.00	80.00	34.00	0.25	-0.06

Observing the summary statistics table, one can find that the respondents learn about climate change from their teachers, parents, friends, others, news or the movies. Striking is the fact that the respondents seem to learn the least from their parents, with the lowest average score being 3.97. Moving on to the variables that gauge how trustworthy the respondents find different sources, with politicians, rated most untrustworthy, with a score of 2.29 and scientists most trustworthy, with a score of 6.01. On average the young people rate themselves as being better in the English language than mathematics, are willing to spend money to fight climate change, even if others do not. On average, the respondents think that around 54% of people in their country think climate change is a serious problem, and that 35% of people in their country would spend money to tackle climate change. Furthermore, the youngsters see themselves as relatively knowledgeable on the topic of climate change, are willing to give up certain things now to benefit from it in the future, are rather risk-seeking and willing to help others, even when nothing is done in return. The average score rated for the respondents' family's wealth is in the middle. The young people see themselves positively showing high levels of agreement on never hiding mistakes, taking things not belonging to them, gossiping about other people's business, obeying laws even when it is unlikely they will be caught and avoid eavesdropping. Finally, youngsters, on average, disagree with the fact that men have priority over women when jobs are scarce.

Factor variables

Moving on to the histograms of the factor variables, one can observe the following: Figure 8 in the appendix shows that most of the respondents are around the age of 16, 17 and 18 (left) and that there are relatively more females in the dataset (right). Figure 9 shows that most of the respondents have enjoyed a university education (left), relatively speaking. Figure 10 in the appendix shows that the largest group is that of 20+ experienced days of extreme weather in the last year and that most of the respondents expect to experience more extreme weather days in the year to come, relatively speaking (Figure 11). Figure 12 shows that most of the youngsters agree with the fact that they work hard most of the time (left), but disagree that they eat more than they should. Furthermore, practically all respondents like visiting new places, observed in Figure 13 (left), and most youngsters disagree that advertisements are silly (right). Figure 14 in the appendix shows that more than 60% of the respondents' days follow a routine (left).

Attitude-behaviour gap by demographics and country-level data

Through examining histograms of the dependent variable *Gap*, by certain groups, interesting patterns could become apparent. One can find the attitude-behaviour gap by gender in Figure 15 (see appendix), where no clear distinction can be observed. Those respondents that answered *Other* more often show a gap in their attitudes with regards to their sustainable behaviour, relatively speaking. The chi-squared test of independence shows that there is no difference in the gap classes based on gender, $\chi^2(2, N = 3.440) = 2.16, p = .34$. Observing Figure 16 in the appendix, showing the classes by age categories, one again finds no clear distinction between age categories with respect to the attitude-behaviour gap, which is confirmed by the chi-squared test, $\chi^2(12, N = 3.440) = 12.28, p = .42$. The most interesting is the final figure in the appendix, where one can clearly see that those respondents with no formal education show an attitude-behaviour gap more often, relatively speaking. Additionally, this is the only group that shows a gap more often than not. It is good to note, however, that this group of youngsters, without a formal education, is extremely small. Complementarily, no clear distinction can be observed with the bare eye, and once again the independence test confirm there to be no association between the gap classes and education, $\chi^2(6, N = 3.440) = 7.82, p = .25$. Finally, the external variables merged on the country level are examined. Starting off with the score capturing overall water risk in a country: in Figure 2 (left) below, one can find the boxplots by class. As one can see the risk score seems to be higher for the *NoGap* class on average. The Welch two-sample t-test confirms this, and rejects the null hypothesis, implying that the *Gap* class show higher levels of *WaterRiskScore2019* than the *NoGap* class, $t(2461.5) = -9.5, p < 2.2e - 16$. Secondly, the SDG index score over 2019 is examined (see boxplot in figure below) to uncover whether the mean SDG score differs per class. Again, the null hypothesis is rejected, meaning that the *Gap* class shows higher levels of *SDGIndexRank2019* than the *NoGap* class, $t(2067.3) = 2.2, p = .03$.

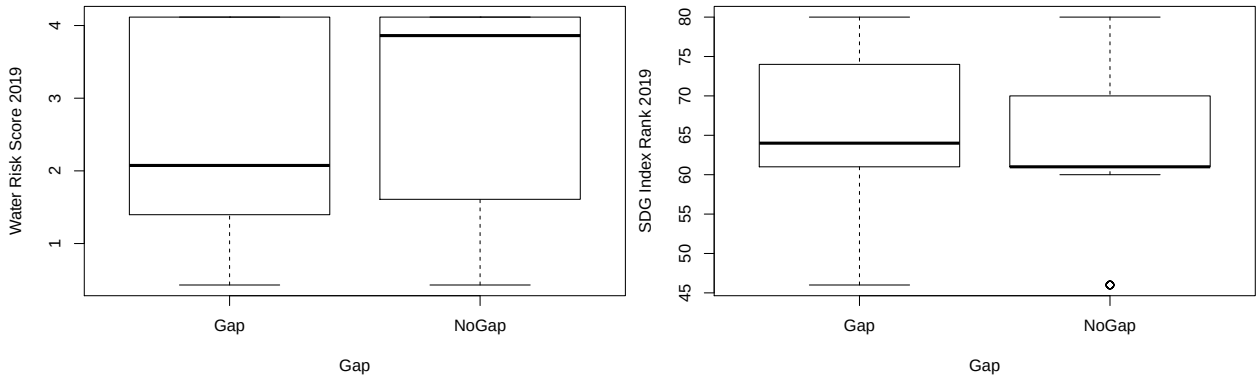

Figure 2: Boxplots of variables WaterRiskScore2019 by class (left) and SDGIndex2019 by class (right)

Methodology

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised learning, dimension-reduction technique, that constructs components or dimensions, as latent variables, to which the original variables relate imperfectly; by establishing underlying dimensions through reducing the data, one can better observe underlying relationships. This non-parametric technique finds a way to reduce dimensions of a high-dimensional space, into a low-dimensional representation of the data, while explaining most of the variance and, therefore, information (James, Witten, Hastie and Tibshirani, 2013). Intuitively speaking, through shared variance between variables, the technique is able to construct components that acknowledge this association between the variables with shared variance.

The resulting components of PCA are computed through the original variables using linear equations; it is good to note that the components found through rotation are linearly uncorrelated (orthogonal), to overcome capturing too much linear variance. Thus, the first component is the standardised linear combination of the features accounting for the largest variance. The second dimension is then generated considering the second-largest explained variance, under the condition that this component is orthogonal, and thus uncorrelated with regards to the first component. This process is repeated for all features, implying PCA can maximally obtain p components equal to the number of variables considered. These original variables can be seen as manifest variables, which PCA tries to explain through generating factors that account for the variance observed in these manifest variables. Ideally, PCA establishes a solution, the rotation matrix, in which some of the manifest variables' component loadings are high for some factor, while keeping the loadings of other variables low for that component; this helps to interpret every factor through a subset of the manifest variables.

A good example of how rotation can enhance interpretation is given by Chapman and Feit (2019) which use the analogy of a pizza. This pizza has different toppings and needs to be cut in a specified number of slices, through different rotations, the resulting possibilities of different slices are very large but are equivalent in terms of mathematics since together they create distinct slices of the underlying basis. Some rotations, however, are more useful since one can better categorise them into slices with topping A, topping B and so on; the underlying basis becomes more interpretable in terms of having differentiated slices, by using different rotations. Equivalently, PCA approaches a matrix of variables the same way, and through different ways of rotating, resulting in increased interpretable components. PCA uses matrix algebra to compute its components. The method uses correlations between the variables to compute the desired factors. Classically,

the obtained model through PCA can be mathematically explained as follows, a matrix $\mathbf{X}$ consisting of p variables (X_1, X_2, ..., X_p), can be explained by m underlying factors (F_1, F_2,..., F_m). X_j is manifest variable j, represented in the underlying latent component. The method, therefore, assumes m underlying dimensions of which the manifest or observed variables are linear functions of these dimensions, with a residual which is unexplained, similar to an error. The variables are, therefore, described by the following linear mathematical equation (Yong and Pearce, 2013):

$$X_j = a_{j1}F_1 + a_{j2}F_2 + ...a_{jm}F_m + e_j \tag{1}$$

with a_{j1}, a_{j2}, a_{jm} the component loadings of variable j on the respective component. The obtained component loadings, therefore, give an idea of how much the variable contributes to a certain factor, which can be interpreted as a weight, with a larger (absolute) weight showing larger importance.

Viewing it from the other perspective, the first factor of a number of variables $\mathbf{X}$ is the normalised linear combination of these features that explain the largest variance, represented mathematically by the formula below (James et al., 2013):

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ...\phi_{p1}X_p \tag{2}$$

where $j = 1, 2, ...p$. Using standardised variables the algorithm searches for that linear combination of variables in the form as above, that explains the most variance, while being subject to the constraint that the sum of the squared loadings is 1. In other words, to find the first dimension of PCA the following optimisation problem is solved for ϕ_{11},... ϕ_{p1}, namely (James et al., 2013):

maximise

$$\frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{j1}x_{ij})^2,$$

subject to

$$\sum_{j=1}^{p}\phi_{j1}^2 = 1 \tag{3}$$

Assuming that the variables in $\mathbf{X}$ are standardised to be z-scores. The standardisation of the features is important, since reducing dimensions using PCA on unscaled variables, can potentially obtain a rather large variance due to large and differing scales between the original variables. The maximisation problem shown above in Equation 3 can be solved using a linear algebra method, called eigendecomposition. Using this

technique the first principal component is found by fitting the best line that captures most of the variance, obtaining a vector of loading scores called the Eigenvector of this component. Additionally, the sum of squared distances for the best fit line with regards to the data points of the first principal component is the Eigenvalue λ^2, which is the same as $var(z_s)$ for component s. After obtaining the Eigenvalues of all components, one can utilise these values to compute the proportion of variance each component explains (PVE), a measure of dimension importance, which comes down to dividing the Eigenvalue λ_s of dimension s by the total number of dimensions, which is equivalent to p, the number of variables used in PCA (James et al., 2013).

PCA uses different rotations to increase interpretability of the components, aiming to load the manifest variables on as few factors as possible with a high loading. Rotation wise there are two broad categories, namely the orthogonal rotation and oblique rotation. The former rotates with an angle of 90° relative, which implies zero correlation between the factors (DeCoster, 1998). The other category concerns the oblique rotation, where the factors are not being rotated by a 90° angle, implying room for some correlation between the factors. Common rotations of the orthogonal category are considered to be Quartimax and Varimax and for the oblique category, Direct Oblimin and Promax (Yong and Pearce, 2013).

One of the most important questions in search of interpretable components is determining how many of these components to obtain. In theory, one can obtain as many dimensions as there are variables, explaining 100% of the variance, but this is not the desired outcome, since the aim of PCA is dimension-reduction, into interpretable, descriptive categories, representing a subset of the manifest variables. Implementing PCA, various diagnostics are performed. Several approaches can be used to decide on the number of principal components. The most common way of determining the number of dimensions is by using a scree plot, which plots the PVE, or Eigenvalues λ_i of the respective dimensions. The scree test works through two conditions: firstly, the eigenvalue of a specific factor needs to be at least 1, which is Kaiser's criterion (Kaiser, 1960). Secondly, the scree plot, sometimes called the elbow plot, should show a point of inflexion (Yong and Pearce, 2013) similar like that of an elbow, where the drop in explained variance, or Eigenvalue of that dimension, is smaller, relatively than the jumps before. PCA can be limited in its interpretation, through that some names of factors may not be easily interpretable in terms of the variables. Secondly, the interpretation might be hard for some factors, due to the fact of split component loadings of a variable, which correlates to both dimensions (Yong and Pearce, 2013).

In the 2-dimensional space, one can use two indicators, namely direction and angle, to help interpret underlying relationships; the closer, angle wise, arrows are, the higher the association, where the direction determines the relative direction (positive or negative relationship) (Chapman and Feit, 2019). However, a simple way of interpreting the components is just by looking at the loadings, which are a measure of the strength of the relationship between the manifest variables and the respective dimension. Thus, interpretation can be derived by identifying the biggest loadings, while also considering the low or zero loadings to confirm the identified factors. Furthermore, one can also use a cut-off point of which loadings are considered to be too low to contribute to the meaning of a component, simplifying interpretation.

Elastic Net Logistic Regression

This thesis explores which factors influence the attitude-behaviour gap using two methods to model the relationship between the factors and the gap, using the obtained factors resulting from PCA, which reduces the dimensions, uncovering underlying structures and relationships. Starting with a white-box method, the logistic regression, which often is labelled comprehensive, intuitive and interpretable, whereafter a more black-box method, random forest, will be implemented to evaluate and compare the different methods in terms of interpretability, ease of use and intuitiveness while exploring influential factors with regards to the attitude-behaviour gap.

Firstly, the relationship is modelled using a logistic regression, extended with a regularization technique called elastic net, which allows for feature selection using penalisation. The logistic regression is a highly interpretable, easy to use and intuitive method often used to tackle classification problems. Comparing it to a more advanced machine learning technique, the logistic regression can often be easily outperformed, but it has the advantage of interpretability, where black-box methods often perform better, at the cost of losing this interpretability. The logistic regression is a classical method, which models the relationship of the independent variables to the binary dependent variable. Although one could utilise a linear regression to solve binary classification problems, it is found to be more ideal to apply logistic regression; it is more suitable for these problems due to it being bounded by $[0, 1]$, enabling a more comprehensive interpretation as the likelihood of belonging to either one of the classes of the binary classification problem. In the context of this thesis $p(X) = Pr(Y = 1|\mathbf{X})$ translates to the likelihood of having a pro-environmental attitude, but not behaving in line with this attitude ($p > 0.5$), i.e. attitude-behaviour gap. As mentioned above, preferred is to use a logistic function to model the binary classification problem, shown below (James et al., 2013):

$$p(X) = \frac{\epsilon^{\beta_0+\beta_1 X}}{1 + \epsilon^{\beta_0+\beta_1 X}}, \tag{4}$$

The logistic function shown above is bounded by a $[0, 1]$-interval, implying an S-curved outcome, with probabilities asymptotic to 0 and 1, never fully reaching these values (James et al., 2013). To make the formula above more interpretable one can manipulate the formula into the following:

$$\frac{p(X)}{1 - p(X)} = \epsilon^{\beta_0 + \beta_1 X}, \tag{5}$$

The left-hand side of the formula above is called the odds and can have values between 0 and ∞. The former is related to a very low probability of in this case showing a gap, and the latter to a very high probability of showing a gap. Intuitively, one can explain the odds as follows. If, for example, the probability of portraying an attitude-behaviour gap is $p(X) = 0.6$, so three out of five have a gap, the odds would come down to $\frac{0.6}{(1-0.6)} = 1.5$, which would mean that the chance of having a gap is 1.5 times as likely as not showing a gap. Taking it further, the natural logarithm can be taken of this odds ratio, resulting in the log-odds, which shows that the log-odds are linear in X. This means that when X is increased by a one-unit change, the log-odds changes by β_1. Complementarily, it is good to note that since the relationship between $p(X)$ and X is not linear, β_1 will not increase steadily, but depends on the value of X at that moment. However, similar to the coefficients obtained through linear regression, if the coefficient is positive (negative) it will result in an increase (decrease) in $p(X)$ (James et al., 2013). The coefficients of the independent variables determining the relationship to the dependent variable are obtained through maximum likelihood (James et al., 2013). This method aims to approximate those values of the independent variables β_j, so that the observed gap status of every observation corresponds to the predicted probability $\hat{p}(x_i)$ of portraying an attitude-behaviour gap for the respective observations maximally. Mathematically this comes down to maximising the following likelihood function, found below (James et al., 2013):

$$l(\beta_0, \beta_1) = \prod_{i:y_1=1} p(x_i) \prod_{i':y_i'=0} (1 - p(x_i')), \tag{6}$$

Complementarily, the classical classification method is extended with a regularization technique called elastic net. This technique helps with feature selection through penalisation, shrinking the coefficients of unimportant variables to zero. This method is a combination of two shrinking methods, namely the ridge regression and the LASSO. These methods utilise a tuning parameter λ which controls for how strong the shrinkage should be. Below one can find the elastic net integrated into the logistic loss function (Zou and Hastie, 2005):

$$\hat{\beta}_{Elastic} = argmin_\beta \left[\sum_{i=1}^{n} \{ y_i \ln p(x_{ij}) + (1 - y_i) \ln(1 - p(x_{ij})) \} + \lambda(\alpha \sum_{j=1}^{p} |\beta_j| + (1 - \alpha) \sum_{j=1}^{p} \beta_j^2) \right], \tag{7}$$

with

- β_j : weights to be found for variable $j = 1, ..., p$
- $\mathbf{X}$: matrix of independent variables made up of elements x_{ij}
- y_i : response variable for observation $i = 1, ...n$
- λ : penalty parameter obtained using k-fold cross validation
- $0 \leq \alpha \leq 1$: elastic net parameter, combination of LASSO and ridge.

The overall objective is to minimise the loss function above. Elastic net regularization is a combination of the LASSO ($\alpha = 1$) and ridge regression ($\alpha = 0$). This alpha affects the flexibility of the model, with the LASSO selecting variables, where the ridge regression merely shrinks some coefficients towards zero; a critique on the LASSO is that its feature selection property depends too largely on the data at hand, implying that the regularization technique results in unstable outcomes, for which the elastic net overcomes this issue, by combining the LASSO and ridge; by choosing $\alpha = 0.5$, equal weight is put on both penalty terms. Firstly, the optimal level of λ needs to be obtained, using K-fold cross-validation. This method divides the training data randomly into k number of equally sized folds (for the size of the dataset used, $k = 5$ is chosen, to guarantee having enough observations to estimate the parameter). This method establishes a value of λ that predicts unseen data well by going through a grid of different values of λ; each time, one fold is used to validate the data on, as the test set, using the other $k - 1$, four, folds, for training. The process is performed k times so that every observation has included in the test set once, for every value of the shrinking parameter λ present in the grid. The method aims to minimise the classification error, thus classifying someone as showing a gap, while in reality, this person does not have a gap between their attitude and behaviour; this is done by minimising the error between the actual class ${y_i}^{test(k)}$ and the predicted class $\hat{y_i}^{test(k)}$, resulting in the best performing model. The classification error can be plotted for different levels of the shrinking parameter λ, of which two values are valuable to look at. The first value is λ_{min}, which is that value of λ that minimises the classification error, and the other is λ_{1sd}, which is one standard deviation away from the former. λ_{1sd} obtains a more parsimonious model, by shrinking more unimportant variables to zero. Hence, it being called $\lambda_{conservative}$.

Random Forest

To understand the random forest algorithm one must first understand how decision trees are built; the element forest in the name refers to the collection of (different) decision trees. Decision trees can be used for both regression and classification problems and are very flexible, can deal with high-dimensionality and missing values well (James et al., 2013). One can think of a decision tree as a literal 'decision flow chart', starting at the root, following all the (internal) nodes, to one of the leaves. The splits can be based both on factor variables and numerical variables. Decision trees are rather simplistic and easy to understand due to their intuitiveness (James et al., 2013). However, the downside of this is that a decision tree often cannot compete with more complex machine learning methods. Random forests, on the other hand, are competitive due to their ensembling nature of decision trees, resulting in large improvements in prediction performance, compensated by losing interpretation (James et al., 2013), moving more towards black-box on the spectrum.

Classification trees are constructed using recursive binary splitting. In the end, the tree classifies an observation to belong to a class, based on its most commonly occurring class in that leaf. The classification error rate, which is the ratio of observations that are not part of the most common class, can be mathematically represented by:

$$E = 1 - max_k(\hat{p}_{mk}) \tag{8}$$

This rate is considered to not be sensitive enough for constructing the decision tree, which is where other alternatives are prefered (James et al., 2013).

Most datasets are not perfectly separable, meaning that an observation can have the same values or levels for the features but belong to another class. Starting at the root node, this node receives all rows of the dataset as input, which are split based on some feature. In response to the splitting variable, the data is partitioned into two subsets, which then become the inputs for the obtained child nodes. The goal of the question asked by the splitting criteria is to create the purest possible distribution of the labels or classes. This is continued, until either the child node is unmixed, called a pure leaf, or when continued splitting does not result in a better possible outcome. The trick to building the best tree is to evaluate which questions to ask, when. This can be done using a measure for the impurity, called the Gini impurity (James et al., 2013). This measure quantifies the level of uncertainty at a specific node. To evaluate whether to split further one can quantify how much splitting based on another feature, reduces this uncertainty, measured by information gain. This process is continued until there are no further questions to be asked, or when there's no improvement in purity, which can be evaluated through the information gain. Each node takes a list of rows as input, resulting from the previous split, which is used to establish which question should be asked next. By iterating over the

values of the different features a list of candidate questions is generated. The questions are then evaluated for every observation, with observations being either true or false on that question. The best question to ask is that question that split the observation in such a way that uncertainty reduction is maximised. Thus, Gini impurity gives the status quo of uncertainty in a certain node, with information gain, showing how much uncertainty is reduced by asking one of the questions mentioned above. Starting off with the impurity, this measure is bounded by 0 and 1, representing no and high uncertainty, respectively. The mathematical equation for the Gini index, can be found below (James et al., 2013):

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{9}$$

This index is small when $\hat{p}_{mk}$s are either close to 0 or 1. This implies, as explained above, that impurity is low when a node consists of mainly observations belonging to one class. In words, the impurity is the sum of the products of the probabilities of classifying every class correctly and incorrectly. The information gain is calculated as follows, one first calculates the impurity of the node from which the new nodes result. Then the weighted average impurity is calculated for the resulting two nodes. By subtracting this average impurity of the child nodes from the parent node, the information gain is calculated. By maximising this gain one can obtain the most optimal question to ask next. An alternative measure of uncertainty is the entropy, similar to Gini numerically speaking (James et al., 2013) and used in the same way to evaluate splits. Entropy is represented mathematically as follows (James et al., 2013):

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk} \tag{10}$$

A downside of a simple decision tree is that a tree is at risk of overfitting, resulting in low-generalisability and suffers from high variance, meaning that a tree can be rather non-robust in the sense that small changes in the dataset can alter the tree significantly (James et al., 2013). To overcome this downside of decision trees, one can create a random forest, which is a collection of different decision trees. These decision trees are constructed the same way as normal but have two additional constraints or rules, namely bootstrapping and randomisation of selected features. To gain a deeper understanding of the random forest algorithm, let's first dive into the bootstrapping part, called bagging, short for bootstrap aggregation (James et al., 2013). This technique, introduced by Breiman (1996a), can be used to reduce the variance of a particular statistical method. The bootstrapping principle finds its value in averaging a set of observations, which in turn decreases variance (James et al., 2013); what this technique does is that it predicts different models using different training sets and averages out the predictions, making the outcome more stable. Thus, mathematically

different models are generated $\hat{f}_1(x), \hat{f}_2(x), ..., \hat{f}_B(x)$ with B training sets, which are averaged to establish a lower-variance model, represented by the following formula (James et al., 2013):

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x) \tag{11}$$

Realistically speaking there is only one training set, but by bootstrapping samples from this training set randomly, one can create different training sets, coming down to the same equation, but using the bootstrapped training sets, given by (James et al., 2013):

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \tag{12}$$

For classification problems the simplest way of averaging the predictions is to evaluate for a single observation the majority vote classification of the models, meaning that the most common label will determine the class. A result of bagging is the out-of-bag error estimate, which has proven to be a good estimate of the test error of such a model, making the necessity of a test set obsolete. It can be shown that roughly every bagged tree uses approximately 63% of the observations, meaning that the remaining 37% observations are not used to fit the bagged model (Breiman, 1996b). These observations are called out-of-bag (OOB) observations (James et al., 2013). These observations can be used to compute the OOB accuracy or classification error, which corresponds highly to that accuracy computed using a separate test set (Gislason, Benediktsson and Sveinsson, 2006; James et al., 2013). Bagging improves accuracy (lowering variance) making the model more competitive, but also comes at the cost of interpretation, making it hard to find which variables are important, normally established using the Gini Index for classification trees. For all bagged trees the Gini indices are summed up and averaged over the B trees (James et al., 2013). The downfall of bagged trees is that the trees are highly correlated with each other, due to the fact that the most important variable, represented by the Gini index, will often be chosen first. This high correlation between the trees can lead to overfitting, which is where the random part of the random forest algorithm comes in. To decorrelate the trees at every split m predictors are considered as candidates from the full set. M is often chosen as the $\sqrt{p}$. Thus, by decorrelating the trees the variance is even further decreased. By only considering an m amount of predictors, this gives chance to weaker predictors. Thus, bagging and random forest are equivalents when $m = p$ (James et al., 2013).

Confusion matrix

The models' performances will be evaluated using a confusion matrix. This matrix shows how many observations have been predicted both correctly and incorrectly. In the case of a binary classification problem, the confusion matrix is made up of four elements, two rows and two columns. The columns and rows represent the predicted and actual class, respectively. When an observation is predicted as having *NoGap*, and this is the actual class, then the observation goes to the top-left entry, being a true positive (TP). In the case that the actual class was the *Gap* class, the observation moves to the bottom-left element, representing false negative (FN), a Type II error (Tharwat, 2018). The same applies for the other class, in this case the *Gap* class. When the predicted class is correct, it moves to the bottom-right element, representing the true negatives (TN), and to the top-right, false positives (FP), a Type I error, also seen as a false alarm (Tharwat, 2018), when not predicted correctly (Luque, Carrasco, Martín and de las Heras, 2019).

The confusion matrix uses these elements to compute metrics, which are evaluated for the models' performance. Firstly, the overall accuracy is computed in the following way (Luque et al., 2019):

$$Accuracy : \frac{TP + TN}{TP + FN + TN + FP} \tag{13}$$

In words, the accuracy are those observations predicted correctly divided by all observations.

However, it is interesting to zoom in on how well each separate class is predicted by the models. Thus, the sensitivity metric computes the accuracy of predicting the true positive class correctly, mathematically represented by (Luque et al., 2019);

$$Sensitivity : \frac{TP}{TP + FN} \tag{14}$$

On the other hand, one might want to know how well the true negative class is predicted, represented by the specificity (Luque et al., 2019:

$$Specificity : \frac{TN}{TN + FP} \tag{15}$$

Finally, the accuracy might be more optimistic than is the case, since classes can be highly imbalanced, meaning that accuracy can come close to 95% if the underrepresented class is very small, not affecting the prediction performance, due to all observations being predicted into one class. Thus, the balanced accuracy is a more realistic measure, represented by (Tharwat, 2018):

$$BalancedAccuracy : \frac{1}{2}(Sensitivity + Specificity) \tag{16}$$

Imbalanced classification problem

Before splitting the dataset into training and test set for further analysis, one should consider whether the problem at hand is a balanced classification problem, meaning that the proportions of both of the classes should be (close to) equal. Unfortunately, this is not the case for the attitude-behaviour gap on sustainable water consumption, with 35% of respondents showing a gap, and 65% showing no gap. For the models not to be influenced too heavily by the overrepresented class, and to balance out the model's performance of predicting both classes, the training set has been manipulated to become a balanced binary classification problem. Chosen is to make use of the *ROSE* package to oversample the minority class, whilst undersampling the majority class, to balance out this manipulation, resulting in a balanced binary classification problem. It is important to first split the data into training and test set and to only correct for class imbalances in the training dataset, to overcome having an oversampled observation in the test dataset; this would result in seemingly better performing models, with higher accuracies. 80% of the data has been used for training, leaving 20% of the data to validate the models' performances. The random forest algorithm can deal with many different formats of variables, whilst the elastic net logistic regression prefers one-hot encoded matrices in the presence of factor variables.

Results

PCA

The first step in the analysis of this research embarks with reducing the dimensions by performing principal component analysis. Before running the function to obtain the principal components, one has to evaluate how many components should be extracted, either using the *Kaiser's* criterion ($\lambda > 1$) or by looking at the scree plot. It is good to note that before performing the analysis, the variables must be normalised, so that unusually large scales do not capture all variance, skewing the results. In the table below, one can find the *eigenvalues* of the respective principal components in the first row, with the proportion of variance and cumulative variance explained per components below. One can see that up until the 8th component the λ is larger than 1 and that the first 8 components explain 54% of the total variance approximately.

Table 2: Eigenvalues and (cumulative) variance of respective components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.925037	1.723651	1.567001	1.323858	1.178468	1.12706	1.116259	1.03919
Proportion of Variance	0.127790	0.102450	0.084670	0.060430	0.047890	0.04380	0.042970	0.03724
Cumulative Proportion	0.127790	0.230230	0.314900	0.375340	0.423230	0.46703	0.510000	0.54724

To make it easier to establish how many components to extract, one can observe the scree plot below in Figure 3. The scree plot helps to determine how many dimensions to reduce the data to, by finding an elbow, or inflexion point. One can see that the inflexion point is situated at the 5th component, whereafter the proportion of explained variance levels out largely. Therefore, it is chosen to obtain five principal components, which explain approximately 42% of the variance.

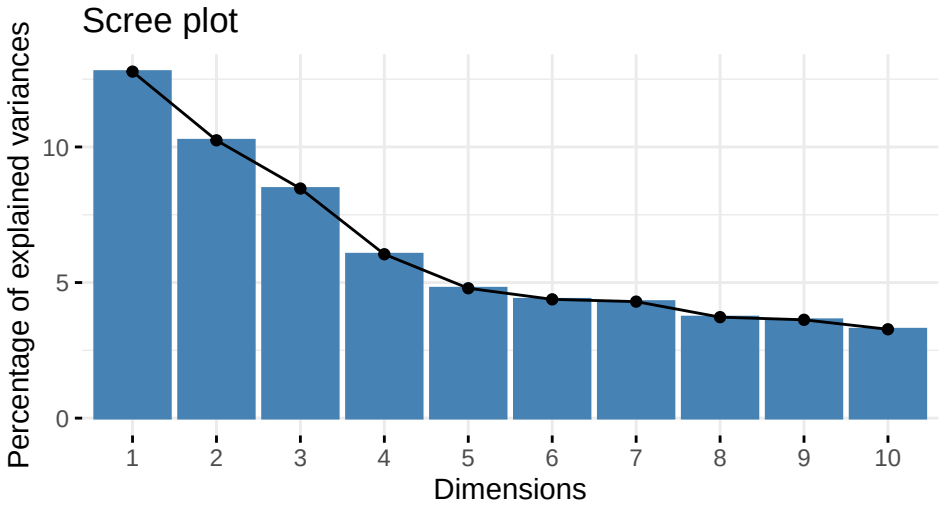

Figure 3: Scree plot of components obtained through principal component analysis

Now that the number of dimensions are established, one can obtain these principal components. The resulting component loadings load onto the respective factors. To enhance interpretation the raw loadings can be rotated using various rotations. Chosen is the *varimax* rotation technique, which results in the most interpretable dimensions. In the table below, one can find the component loadings of the variables to the components. In name of clarity the loadings have been cut off ($r > 0.3$), resulting in variables loading sufficiently high onto a dimension, making interpretation easier. Firstly, the first component shows that the variables *EnglishLanguage*, *WillingSpendMoneyFightCC*, *WillingSpendMoneyFightCCEvenIfOthersDont*, *RelativeKnowledgeCC*, *WillingnessGiveUpNowBenefitFuture*, *WillingnessTakeRisks* and *WillingnessHelpOthersWithoutAnythingInReturn* load highly onto the dimension. Taking these variables one could interpret this dimension as one that captures altruism, relative knowledge on climate change and future orientation, which go hand in hand very well.

Table 3: Component loadings of variables on respective dimensions

	PC1	PC2	PC3	PC4	PC5
LearnFromTeachers		0.34			-0.44
LearnFromParents					-0.50
LearnFromFriends					-0.63
LearnFromOthers					-0.40
LearnFromNews		0.50			
LearnFromMovies		0.44			
TrustPoliticians		0.56			
TrustMedia		0.74			
TrustScientists		0.36			
TrustCelebrities		0.65			
MathematicsSkill					
EnglishLanguage	-0.41				
WillingSpendMoneyFightCC	-0.70				
WillingSpendMoneyFightCCEvenIfOthersDont	-0.72				
ShareYoungPplThinkCCSeriousProblem					-0.75
ShareYoungPplSpendMoneyTackleCC					-0.72
RelativeKnowledgeCC	-0.50				
WillingnessGiveUpNowBenefitFuture	-0.66				
WillingnessToTakeRisks	-0.63				
WillingsnessHelpOthersWithoutAnythingInReturn	-0.55				
FamilyEnoughMoneyToBuyWhatIWant				-0.85	
FamilyDontHaveWorriesPayingBills				-0.89	
ExpectFamilyNoMoneyWorryFuture				-0.80	
NeverHideMistakes			-0.61		
NeverTakeThingsDontBelongToMe			-0.58		
DontGossipOtherPeoplsBusiness			-0.70		
AlwaysObeyLawsEvenUnlikelyToBeCaught			-0.54		
AvoidEavesDroppingPrivateConversations			-0.67		
ScarceJobsMenPriority		0.34			

Moving on to the second component, one can observe that the sources from which an adolescent learns, and the level of trust in different sources are captured in this dimension, meaning this dimension controls for the level of trust the respective adolescent has in different sources of information on the topic of climate change. Furthermore, the third component is represented by the variables *NeverHideMistakes*, *NeverTakeThingsDontBelongToMe*, *DontGossipOtherPeoplesBusiness*, *AlwaysObeyLawsEvenUnlikelyToBeCaught* and *AvoidEavesDroppingPrivateConversations* which captures the level of social desirability bias of the respondent. The second to last component is very well represented by the variables that are about the wealth status of the family, implying the level of privilege and accessibility to certain resources. Finally, the last component loads highly on different sources of information and the variables on peer perspective with regards to (tackling) climate change, implying that this component can be interpreted as a dimension that captures social norms.

Elastic Net Logistic Regression

The second method used is the elastic net logistic regression, where the logistic regression is extended with a regularization technique, penalising unimportant variables, performing variable selection. Using 5-fold cross-validation one can obtain the levels of λ that minimise the misclassification error and is one standard deviation away from this minimum error, λ_{min} and λ_{1sdev} respectively. The misclassification error plot can be found below. Figure 4 shows how many variables are selected at both values of λ. The conservative model, using λ_{1sdev} excludes too many variables, considering the dimensions have already been reduced significantly performing PCA. Therefore, the model using λ_{min} is preferred.

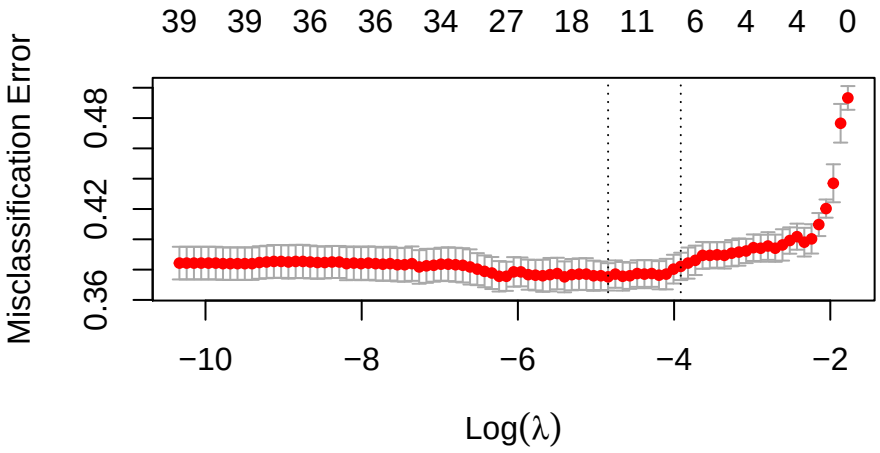

Figure 4: Misclassification error plot for different values of lambda

In Table 4 below one can find the final model. This model contains all the five principal components and the variables *WaterRiskScore2019*, *Experienced11to15Days*, *Age24*, *ExpectedFewerDays*, *EducationToAge12*, *SDGIndexRank2019* and *Age19*. The coefficients are sorted in descending order, showing that *WaterRiskScore2019*, _Altruism, *FamilyWealth* and *Experienced11to15Days* and *SocialDesirabilityBias* are the five most important variables. As established in the section above, the most important variable, *WaterRiskScore2019* controls for the degree at which the respondents' living environment is at risk of water scarcity. The variable shows a negative sign, meaning that the higher the score the lower the probability of portraying a gap between attitude and behaviour. The second-largest coefficient is that of the variable *Altruism* which represents the level of future orientation and altruism of the respondent. The sign of this coefficient is positive, meaning that the probability of portraying a gap increases in this component. Furthermore, the third-largest coefficient is that of variable *FamilyWealth*, which shows a negative sign, meaning that the likelihood of behaving in line with one's attitude increases in this component. The fourth most important variable is the level *11to15Days* of *ExperiencedExtremeWeatherDays*, with a negative sign shows that for this level the probability of having a gap is lower than the reference level. Interestingly, the fifth most important variable is that component that proxies for the level of social desirability bias. This variable has a positive impact on the gap, meaning that the probability of having an attitude-behaviour gap increases in this component.

Table 4: Optimal model obtained through elastic net logistic regression

	name	coefficient
7	WaterRiskScore2019	-0.29
9	Altruism	0.25
12	FamilyWealth	-0.19
5	Experienced11to15Days	-0.19
11	SocialDesirabilityBias	0.18
3	Age24	0.13
6	ExpectedFewerDays	-0.11
4	EducationToAge12	0.08
8	SDGIndexRank2019	-0.06
1	(Intercept)	-0.04
10	TrustSources	-0.02
2	Age19	-0.02
13	SocialNorms	0.02

Finally, using the obtained model, the prediction performance is determined by evaluating its performance on unseen data, the test set. The resulting accuracies and other metrics can be found in Table 5 below. The model predicts better than chance with an accuracy of 60.84%. However, it is more interesting to look at the balanced accuracy, considering that the problem at hand is of an imbalanced nature. The balanced accuracy is close to the overall accuracy with 59.63%. These accuracies imply that both classes are predicted not equally well, but are predicted similarly well. The final two metrics confirm this, with a sensitivity and specificity of 63.62% and 55.65%, respectively.

Table 5: Confusion matrix of optimal elastic net logistic regression model

	No Gap	Gap	Accuracy	Balanced Accuracy	Sensitivity	Specificity
No Gap	285	106	60.84%	59.63%	63.62%	55.65%
Gap	163	133				

Random Forest

The last model used, a black-box method, is the random forest algorithm. This algorithm is basically a collection of different decision trees. However, to overcome that the most important variables are chosen for the splits, the algorithm decorrelates the trees in the forest by only considering m randomly chosen candidate variables. Before the final model is established, one should tune the random forest. The parameters to be tuned are the m candidate splitting (*mtry*) variables and the size of the forest, represented by *ntree*. Below one can find the grid search for the amount of randomly selected predictors, using 5-fold cross-validation. The grid is iteratively searched for *mtry* from two to nine. In Figure 5 one can find that the optimal amount of splitting variables is four. Literature suggests that the optimal amount is (close to) the $\sqrt{(p)}$, which is four. Thus, chosen is $m = 4$, obtained through the grid search, is in line with the literature.

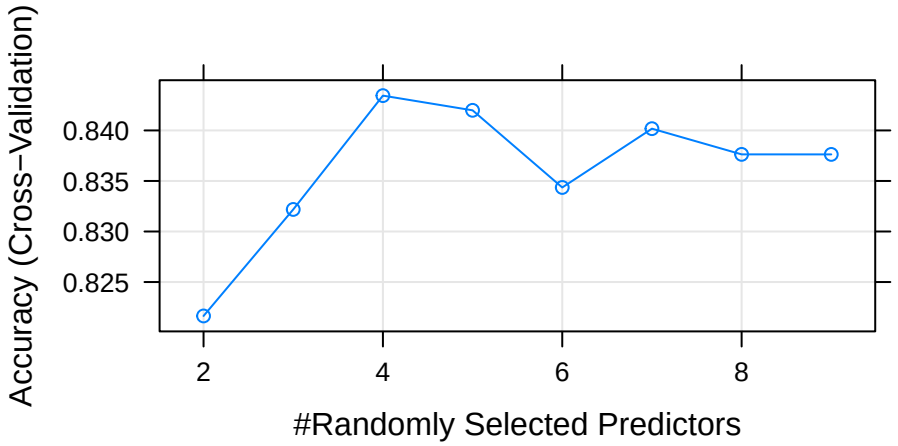

Figure 5: Grid search for mtry

Secondly, the amount of trees the forest contains needs to be established. This can be done by plotting the OOB error rate, which is a good proxy for the error rate of the test set over the number of trees. Normally, more trees are better, but due to its computational inefficiency one chooses for the number of trees at which the error rate stabilises. Observing Figure 6 below, one can see that after $ntree = 300$ it levels out and becomes stable. Therefore, chosen is $ntree = 300$.

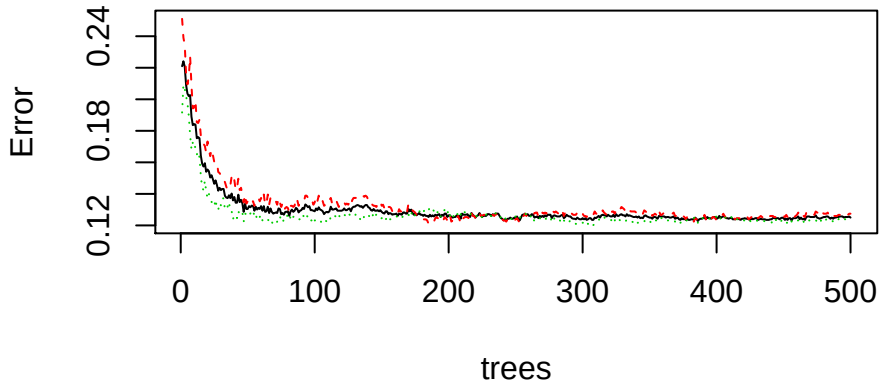

Figure 6: OOB-error for ntree for random forest

Now that the final model is established with $mtry = 5$ and $ntree = 300$, the final model can be established. Black-box methods get their name from decreased interpretability and intuitiveness. Where the logistic regression outputs a model with coefficients, which show the direction of the effect of the variables and the importance through the size of the coefficients (when the variables are normalised), the random forest algorithm merely has a plot for the variable importance (see Figure 7). The mean decrease in Gini impurity averaged over B trees, shows how important a certain variable is. Comparing the five most important variables to those of the logistic regression, the factors obtained through PCA show to be the most important. Firstly, the component that captures the level of altruism shows to be the most important variable, whereafter *SocialDesirabilityBias* comes second. Moreover, one's privilege captured in the component that says something about *FamilyWealth* is the third most important variable. Fourthly, *SocialNorms* comes in fourth place and *TrustSources* is the fifth most important variable. The difference between the random forest algorithm and the elastic net logistic regression is that no variables are excluded in the random forest, due to the absence of a feature selection technique. The other most important variables included in the top 10 after the components are *Age*, *WaterRiskIndex2019*, *ExperiencedDaysExtremeWeatherLastYear*, *HighestEducation*, _SDGIndexRank2019, in that order of descending importance. The last variables show close to 0 importance in terms of the mean decrease in Gini impurity.

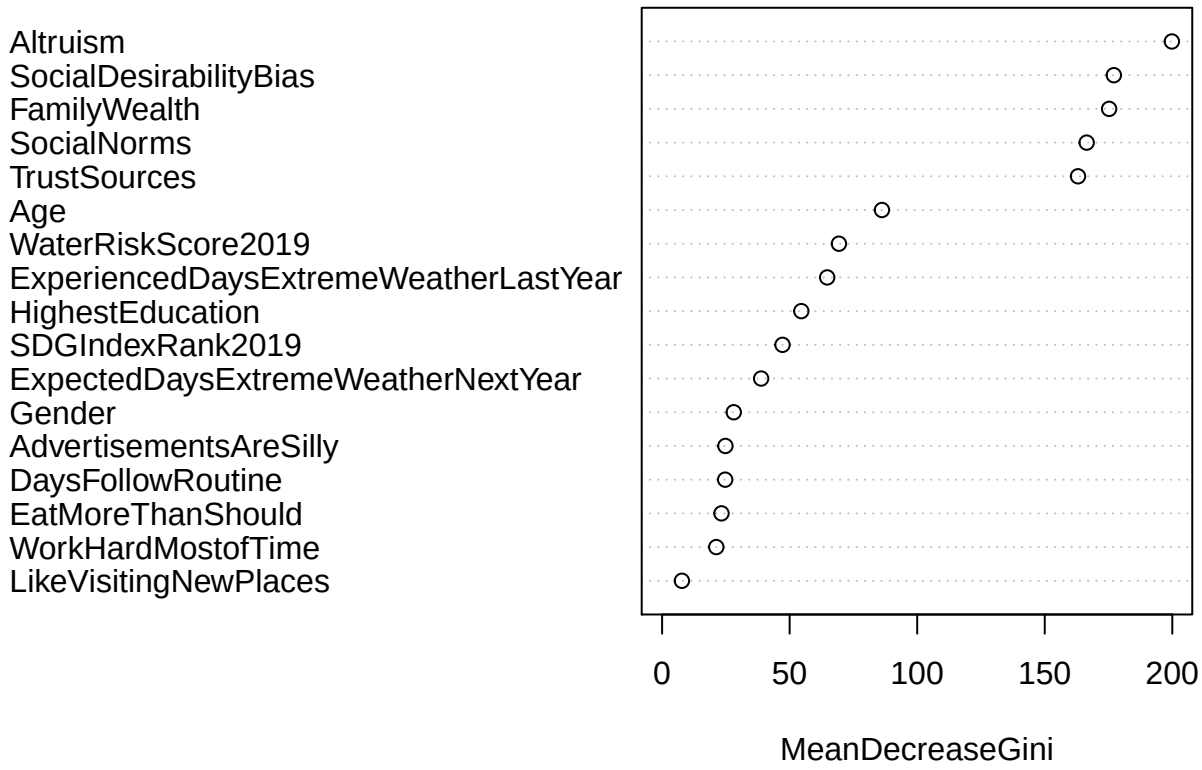

Figure 7: Random Forest Variable Importance

Finally, now that the model has been established and the most important variables have been examined, the model's performance can be established. This is done, equivalent to before, by evaluating this performance on the held apart test set. Below one can find the table showing the confusion matrix, in combination with additional metrics. Straight away, one can see that the accuracy of the random forest model is 62.45% which is just a bit better than the logistic regression model. However, the balanced accuracy is lower with 57.83%, implying that the overrepresented class, namely *NoGap*, is predicted better than the *Gap* class. This is confirmed by the relatively lower specificity of 42.68% and a higher sensitivity of 72.99%, compared to 55.65% and 63.62% for the logistic regression model, respectively.

Table 6: Confusion matrix of random forest model

	No Gap	Gap	Accuracy	Balanced Accuracy	Sensitivity	Specificity
No Gap	327	137	62.45%	57.83%	42.68%	72.99%
Gap	121	102				

Conclusion

This research embarked on examining the attitude-behaviour gap with respect to sustainable water conservation behaviour in adolescents, ranging from age 13 to 24 and older. In collaboration with Facebook, KidsRights and the Erasmus University approximately 11.000 teenagers and young adults have been surveyed regarding a wide variety of questions. This paper focuses on comparing two machine learning methods, namely elastic net logistic regression, which is a more white-box method, due to its interpretability, versus the more black-box method, namely the random forest algorithm. Complementarily, before comparing these methods, the many dimensions available due to the survey are reduced using a dimension-reduction technique. After PCA, the established components are then translated into these dimensions representing the original variables for every observation used in the elastic net logistic regression and random forest model.

Performing PCA, five dimensions have been obtained. As explained in the results section, the first component represents the dimension that says something about one's degree of altruism, relative knowledge on climate change and future orientation. The second component seems to say something about the degree of trust a respondent has in different sources of media and learning from various media outlets, representing more external information and knowledge transfer. The third dimension is represented by those variables that were used to control for social desirability bias. Second to last, the fourth component is represented by those variables that say something about family wealth and level of privilege in terms of not having to worry about finances now or in the future. Finally, the last component resembles sources from which the respondent learned about climate change, in a closer environment, from teachers, parents, friends and others in combination with those variables about the stances of peers on climate change and whether they would spend money to tackle the problem of climate change, which seems to say something about social norms.

After these dimensions have been established, making the original, manifest variables obsolete, and are therefore excluded from further analysis, this research moves on to integrating the obtained dimensions into the models. Compared are the elastic net logistic regression and the random forest algorithm. Both models include all factors obtained through PCA, and rank the *Altruism*, *FamilyWealth* and *SocialDesirabilityBias* factors of similar importance. The logistic regression model shows the water risk index to be the most important variable. This fact can be explained through the construal theory of Trope and Liberman (2010); the higher the risk score of living under water scarcity in a country, decreases the likelihood of portraying a gap between attitude and behaviour with regards to sustainable water consumption, due to smaller psychological distance. On the other hand, the random forest model ranks the factor *SocialNorms* higher relatively speaking, which could be explained by the fact that the models control for the water risk score, which says something

about the water scarcity risk in a country. The variable controlling for water scarcity could reflect the social norms in a country, namely that of sustainable water consumption due to the population living in areas where scarcity in water resources is a characteristic of the living environment. This could mean that the water risk variable pulls the importance of *SocialNorms* down because the effect is already captured in the water risk index and vice versa.

Comparing the performance of the models, merely looking at the accuracy one would favour this model since the overall accuracy is 62.45% here, but when having a closer look it becomes clearer that both classes are not predicted evenly well, resulting in a lower balanced accuracy of 57.83% compared to 59.63% for the logistic regression model. The overrepresented *NoGap* class is predicted better using the random forest model, with a sensitivity of 72.99% compared to 63.62% of the logistic regression, whereas the specificity, measure of how well the underrepresented *Gap* class is predicted is 42.68% compared to 55.65% for the logistic regression. Which model one prefers fully depends on what the person in question aims to achieve. Since this research focuses on which factors seem to affect the attitude-behaviour gap, this paper prefers the more balanced predicting model, which predicts the underrepresented *Gap* class the best. Thus, the elastic net logistic regression is preferred since this class is predicted with a specificity of 55.65% compared to 42.68% of the random forest model. Additionally, the elastic net logistic regression allows for better interpretation through the established model that contains both variables and coefficients, showing the direction and size of their respective effects.

Conclusively, the logistic regression and random forest both suggest that *Altruism*, *FamilyWealth* and *SocialDesirabilityBias* are important predictors of the attitude-behaviour gap. However, the logistic regression additionally finds the variable that controls for risk of water scarcity to be the most important variable, as opposed to it coming after all the factors in the random forest model. Another interesting discrepancy between both is that in the random forest model *SocialNorms* is suggested to be more important than is suggested by the logistic regression model. Furthermore, both models suggest that controlling for *Age*, *Education* and *SDGIndex2019* adds value since these background factors are included in the 10 most important variables in both models. When comparing the performances of the models, the logistic regression model is preferred, due to the fact of its more balanced accuracy, predicting the underrepresented *Gap* class better than the random forest model.

Discussion

Recommendations

Government intervention through policies is important when the individuals in a country behave in a way that does not serve the public good, and one should aim to generate persisting changes in the beliefs and norms held by these consumers so that they alter their behaviour in name of the public good (Kinzig et al., 2013). Nordhaus (2019) adds that for a policy in the context of sustainability to be helpful the consequences of climate change and thus indirectly of mankind's actions need to be understood. Taking the findings that result from the analysis, recommendations can be made to help create more effective policies and education which aim to close the gap between attitude and behaviour

Firstly, both models show the factor *Altruism* to be an important predictor of the attitude-behaviour gap. This factor encapsulates one's willingness to spend money to fight climate change, even if others do not, relative knowledge about climate change, future-orientation and altruism. These variables reflect an awareness or knowledge about the consequences of climate change, which in itself is future-oriented, awareness of responsibility and the fact that one can indeed make a difference. As literature points out, knowledge on the topic and knowledge on the consequences of one's behaviour are important prerequisites for behaviour. Putting that into the context of sustainable water conservation behaviour research has shown that a water footprint increases awareness of one's level of water consumption, in turn increasing environmental awareness (Gómez-Llanos et al., 2020). Complementarily, Haida *et al.* (2019) suggest that, by linking the concept of water footprint to climate change, people will become empowered to adapt their unsustainable water consumption behaviours. Having that said, it is recommended that younger generations get educated on the topic of sustainability and their role, through for example an indicator which uncovers one's footprint on the environment, empowering the youngsters to make choices based on the transparency of the consequences of their behaviour. By making consequences of one's behaviour more salient, increasing awareness of consequences, psychological distance decreases, which in turn increases the likelihood of behaving more sustainable (Trope and Liberman, 2010); the finding that *WaterRiskScore2019* is an important predictor in both models, and the most important predictor in the logistic regression model reflects this, since those that are physically in closer proximity to an environment more at risk of water scarcity, the lower the likelihood of portraying a gap between attitude and behaviour. Complementarily, education or policies should focus on helping the adolescents' perspective to become more focused on the distant future, as Rabinovich *et al.* (2010) suggest to help close the gap between attitude and behaviour; the last component of the SHIFT framework, namely tangibility, highlights the need to make consequences more tangible, by closing the gap of temporal

discrepancy and communicating in a concrete and clear fashion, 'hitting home' by translating impacts to a more local level (White et al., 2019).

Secondly, both models show *FamilyWealth* to be an important predictor of the attitude-behaviour gap, which can be linked to utilitarian beliefs, behavioural control, and hedonic and gain goals. Taking this finding together with the factor that controls for *SocialNorms*, included in both models and shown to be an important predictor in the random forest model, the following recommendations can be made: there is a clear role for government to help form social norms through policy-interventions. As Steg *et al.* (2014) highlight, consumers make decisions by evaluating costs and benefit of one's behaviours. Therefore, they argue that policies should aim to increase and decrease perceived benefits and costs, respectively, by making the pro-environmental behaviour more enjoyable and convenient in terms of time and money. This way the relative importance of normative goals increases compared to hedonic and gain goals, lowering the conflict between them. Additionally, another tactic could be to strengthen normative goals by intervening on the level of biospheric values and beliefs which enhance one's self-identity as being sustainable, in turn increasing the relative importance of the normative goals compared to the hedonic and gain goals. These normative goals can be altered by policy either through highlighting norms, which in turn affects behaviour, or shift social norms by steering on the desired behaviour first (Kinzig et al., 2013). He continues that the former can be achieved by supplying information about both descriptive and injunctive norms. The latter can be achieved by policies that steer on a specific behaviour, for example utilising aggressive tactics such as implementing financial instruments; discontinuation of habits, by forcing formation of new habits, which in turn results in the development of social norms, can be seen as a strategy to create behavioural change toward sustainability (White et al., 2019). Ajzen *et al.* (2011) add that by uncovering one's beliefs with respect to their behaviours, norms and control, behaviour can be altered by leveraging the information at hand. On the one hand, policies should challenge those beliefs that are in favour of the behaviour one wishes to change while providing information to create new beliefs in support of the desired sustainable behaviour. Besides institutions there is a role for the parents to guide their offspring to make better decisions in their day-to-day life; research has shown that indeed adolescents can best be reached through their parents and that parents should be made aware of their role model status. When awareness has been established, parents should show the way by making the pro-environmental behaviour visible toward their children, serving as a norm prevalent in the family environment (Grønhøj and Thøgersen, 2012).

Finally, identification with sustainability goes hand in hand, or even is a prerequisite of the factor *Altruism*; before one is willing to tackle climate change, one must learn about the topic, one must feel drawn to the sustainability debate. For such a topic to become an existent component in one's moral compass, such information must be transferred before, triggering the person in question, by for example knowledge transfers through peers, documentaries or even society-critical movies, implied by *SocialNorms* and *TrustSources*. Thus, as Bartels and Reinders (2016) conclude, whether governmental, educational or organisational, it is recommended to leverage tactics that help to increase identification with individuals that behave sustainably, for example through communicating positive aspects of such consumer groups, increasing positive associations, feelings and cognition (White et al., 2019), establishing an identity that finds sustainability important; Gatersleben *et al.* (2012) confirm the finding that self-identity plays a significant role in portraying sustainable behaviours and suggest that identity campaigns could provide an effective tool to influence sustainable behaviour.

Limitations

The findings of this research are limited in a number of ways: firstly, the survey consists of self-reported behaviour, allowing room for bias to exist since some answers that are given are more socially desirable biasing the answers (Gifford and Nilsson, 2014; de Leeuw et al., 2015) of the respective respondent, which in turn affects the findings of this research. Furthermore, as Ajzen and Fishbein (1977) mention, that in order to predict behaviour using attitude, there needs to be a high correspondence between the two. The gap is now modelled using pro-environmental attitude in combination with the more specific sustainable water consumption behaviour. Thus, the attitude would have to be more specific about sustainable water consumption as opposed to the more general pro-environmental attitude, which can affect the final gap. Therefore, this research is limited in its findings, due to the survey not being designed in the most optimal way, possibly affecting the outcomes. Second to last, the findings of this research are limited in the sense that the externally merged data has been added on the country-level, where India is largely overrepresented, possibly increasing the importance of *WaterRiskScore2019* and *SDGIndexRank2019*, affecting the outcomes of this research. Furthermore, the original dataset started with approximately 11.000 observations, but after processing the dataset, it became clear that only around 3.700 respondents had fully completed the survey, which in turn affects the findings since a larger sample size could possibly have gotten more stable results.

Future research

Firstly, for future research there are a number of things that could be done: firstly, this research should be done over a number of years, following the same observations to see how attitudes, behaviours and the derived gap evolve over time and find which variables affect the gap through time. Secondly, for future research it could be interesting to compare this young sample to different age category samples, to find out whether different variables are of greater or lesser importance between age categories, in turn giving important input to create more effective policies for different target audiences. Thirdly, future research should include indirect water consumption, since this is a large component of overall water usage, often overseen (Gómez-Llanos et al., 2020). Furthermore, besides controlling for many more background factors, mediating and moderating variables should be explored in future research. Finally, future research should survey many more people over many more countries and aim to get an evenly amount of respondents of different countries so better cross-country inferences can be made, making clear whether there are important country-specific factors affecting the phenomenon called the attitude-behaviour gap. It is important to understand the differences between countries, in terms of culture or societal differences, when in the process of designing effective environmental policies (Steg et al., 2014).

References

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. Psychological Bulletin, 84(5), 888–918. https://doi.org/10.1037/0033-2909.84.5.888

Ajzen, I.. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-t

Ajzen, I. (2011). The theory of planned behaviour: Reactions and reflections. Psychology & Health, 26(9), 1113–1127. https://doi.org/10.1080/08870446.2011.613995

Ajzen, I., & Dasgupta, N. (2015). Explicit and Implicit Beliefs, Attitudes, and Intentions. The Sense of Agency, 115–144. https://doi.org/10.1093/acprof:oso/9780190267278.003.0005

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), The handbook of attitudes (pp. 173-221). Mahwah, NJ: Erlbaum.

Bartels, J., & Reinders, M. J. (2016). Consuming apart, together: the role of multiple identities in sustainable behaviour. International Journal of Consumer Studies, 40(4), 444–452. https://doi.org/10.1111/ijcs.12269

Berrang-Ford, L., Biesbroek, R., Ford, J. D., Lesnikowski, A., Tanabe, A., Wang, F. M.,... Heyman, S. J. (2019). Tracking global climate change adaptation among governments. Nature Climate Change, 9(6), 440–449. https://doi.org/10.1038/s41558-019-0490-0

Breiman, L. (1996a). Bagging Predictors. Machine Learning, 24(2), 123–140. https://doi.org/10.1023/a:1018054314350

Breiman, L. (1996b). Out-of-bag estimation. Technical report, 1-13. Department of Statistics, University of California, Berkeley.

Carrington, M. J., Neville, B. A., & Whitwell, G. J. (2014). Lost in translation: Exploring the ethical consumer intention–behavior gap. Journal of Business Research, 67(1), 2759–2767. https://doi.org/10.1016/j.jbusres.2012.09.022

Chapman, C. & Feit, E. M. (2019). R For Marketing Research and Analytics. R for Marketing Research and Analytics, 193–214. https://doi.org/10.1007/978-3-030-14316-9

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. Journal of Personality and Social Psychology, 58(6), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Craig, C. A., Feng, S., & Gilbertz, S. (2019). Water crisis, drought, and climate change in the southeast United States. Land Use Policy, 88, 104110. https://doi.org/10.1016/j.landusepol.2019.104110

de Leeuw, A., Valois, P., Ajzen, I., & Schmidt, P. (2015). Using the theory of planned behavior to identify key beliefs underlying pro-environmental behavior in high-school students: Implications for educational interventions. Journal of Environmental Psychology, 42, 128–138. https://doi.org/10.1016/j.jenvp.2015.03.005

DeCoster, J. (1998). Overview of factor analysis. Retrieved from https://www.researchgate.net/publication/255620387_Overview_of_Factor_Analysis/citations

Duran-Encalada, J. A., Paucar-Caceres, A., Bandala, E. R., & Wright, G. H. (2017). The impact of global climate change on water quantity and quality: A system dynamics approach to the US–Mexican transborder region. European Journal of Operational Research, 256(2), 567–581. https://doi.org/10.1016/j.ejor.2016.06.016

Farrow, K., Grolleau, G., & Ibanez, L. (2017). Social Norms and Pro-environmental Behavior: A Review of the Evidence. Ecological Economics, 140, 1–13. https://doi.org/10.1016/j.ecolecon.2017.04.017

Gatersleben, B., Murtagh, N., & Abrahamse, W. (2012). Values, identity and pro-environmental behaviour. Contemporary Social Science, 9(4), 374–392. https://doi.org/10.1080/21582041.2012.682086

Gifford, R., & Nilsson, A. (2014). Personal and social factors that influence pro-environmental concern and behaviour: A review. International Journal of Psychology, n/a. https://doi.org/10.1002/ijop.12034

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. Pattern Recognition Letters, 27(4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Gómez-Llanos, E., Durán-Barroso, P., & Robina-Ramírez, R. (2020). Analysis of consumer awareness of sustainable water consumption by the water footprint concept. Science of The Total Environment, 721, 137743. https://doi.org/10.1016/j.scitotenv.2020.13774

Gosling, S. N., & Arnell, N. W. (2013). A global assessment of the impact of climate change on water scarcity. Climatic Change, 134(3), 371–385. https://doi.org/10.1007/s10584-013-0853-x

Grønhøj, A., & Thøgersen, J. (2012). Action speaks louder than words: The effect of personal attitudes and family norms on adolescents' pro-environmental behaviour. Journal of Economic Psychology, 33(1), 292–302. https://doi.org/10.1016/j.joep.2011.10.001

Haida, C., Chapagain, A. K., Rauch, W., Riede, M., & Schneider, K. (2019). From water footprint to climate change adaptation: Capacity development with teenagers to save water. Land Use Policy, 80, 456–463.

https://doi.org/10.1016/j.landusepol.2018.02.043

Hoekstra, A. Y., & Wiedmann, T. O. (2014). Humanity's unsustainable environmental footprint. Science, 344(6188), 1114-1117. https://doi.org/10.1126/science.1248365

Hoekstra, A., Chapagain, A., & Zhang, G. (2015). Water Footprints and Sustainable Water Allocation. Sustainability, 8(1), 20. https://doi.org/10.3390/su8010020

Hofste, R. (2019). Aqueduct 3.0: Updated Decision-Relevant Global Water Risk Indicators. Retrieved from https://www.wri.org/publication/aqueduct-30

Howard, G., Calow, R., Macdonald, A., & Bartram, J. (2016). Climate Change and Water and Sanitation: Likely Impacts and Emerging Trends for Action. Annual Review of Environment and Resources, 41(1), 253–276. https://doi.org/10.1146/annurev-environ-110615-085856

IPCC. (2014). Climate Change 2014: Mitigation of Climate Change. Retrieved from https://www.ipcc.ch/site/assets/uploads/2018/02/ipcc_wg3_ar5_full.pdf

IPCC. (2018). Global Warming of 1.5 °C. Retrieved from https://www.ipcc.ch/sr15/

Jaeger, W. K., Amos, A., Bigelow, D. P., Chang, H., Conklin, D. R., Haggerty, R., ... Turner, D. P. (2017). Finding water scarcity amid abundance using human–natural system models. Proceedings of the National Academy of Sciences, 114(45), 11884–11889. https://doi.org/10.1073/pnas.1706847114

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141–151.https://doi.org/10.1177/001316446002000116

Kang, J., Grable, K., Hustvedt, G., & Ahn, M. (2017). Sustainable water consumption: The perspective of Hispanic consumers. Journal of Environmental Psychology, 50, 94–103. https://doi.org/10.1016/j.jenvp.2017.02.005

Kinzig, A. P., Ehrlich, P. R., Alston, L. J., Arrow, K., Barrett, S., Buchman, T. G., ... Saari, D. (2013). Social Norms and Global Environmental Challenges: The Complex Interaction of Behaviors, Values, and Policy. BioScience, 63(3), 164–175. https://doi.org/10.1525/bio.2013.63.3.5

Klöckner, C. A. (2013). A comprehensive model of the psychology of environmental behaviour—A meta-analysis. Global Environmental Change, 23(5), 1028–1038. https://doi.org/10.1016/j.gloenvcha.2013.05.014

Knussen, C., & Yule, F. (2008). "I'm Not in the Habit of Recycling." Environment and Behavior, 40(5), 683–702. https://doi.org/10.1177/0013916507307527

Kollmuss, A., & Agyeman, J. (2002). Mind the Gap: Why do people act environmentally and what are the barriers to pro-environmental behavior? Environmental Education Research, 8(3), 239–260. https://doi.org/10.1080/13504620220145401

Kummu, M., Guillaume, J. H. A., de Moel, H., Eisner, S., Flörke, M., Porkka, M., ... Ward, P. J. (2016). The world's road to water scarcity: shortage and stress in the 20th century and pathways towards sustainability. Scientific Reports, 6(1), 1–16. https://doi.org/10.1038/srep38495

Liu, J., Yang, H., Gosling, S. N., Kummu, M., Flörke, M., Pfister, S., ... Oki, T. (2017). Water scarcity assessments in the past, present, and future. Earth's Future, 5(6), 545–559. https://doi.org/10.1002/2016ef000518

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Mekonnen, M. M., & Hoekstra, A. Y. (2016). Four billion people facing severe water scarcity. Science Advances, 2(2), e1500323. https://doi.org/10.1126/sciadv.1500323

Millington, N., & Scheba, S. (2020). Day Zero and The Infrastructures of Climate Change: Water Governance, Inequality, and Infrastructural Politics in Cape Town's Water Crisis. International Journal of Urban and Regional Research, 1–17. https://doi.org/10.1111/1468-2427.12899

National Public Radio. (2019). Transcript: Greta Thunberg's Speech At The U.N. Climate Action Summit. Retrieved March 16, 2020, from https://www.npr.org/2019/09/23/763452863/transcript-greta-thunbergs-speech-at-the-u-n-climate-action-summit?t=1584360845759

Naumann, G., Alfieri, L., Wyser, K., Mentaschi, L., Betts, R. A., Carrao, H., ... Feyen, L. (2018). Global Changes in Drought Conditions Under Different Levels of Warming. Geophysical Research Letters, 45(7), 3285–3296. https://doi.org/10.1002/2017gl076521

Nordhaus, W. (2019). Climate Change: The Ultimate Challenge for Economics. American Economic Review, 109(6), 1991–2014. https://doi.org/10.1257/aer.109.6.1991

Rabinovich, A., Morton, T., & Postmes, T. (2010). Time perspective and attitude-behaviour consistency in future-oriented behaviours. British Journal of Social Psychology, 49(1), 69–89. https://doi.org/10.1348/014466608x401875

Rosa, E. A., & Dietz, T. (2012). Human drivers of national greenhouse-gas emissions. Nature Climate Change, 2(8), 581–586. https://doi.org/10.1038/nclimate1506

Rosa, L., Chiarelli, D.D., Rulli, M.C., Dell'Angelo, J., & D'Odorico, P. (2020). Global agricultural economic water scarcity. Science Advances, 6(18), eaaz6031. https://doi.org/10.1126/sciadv.aaz6031

Schwartz S. (1977) Normative influences on altruism. In: Berkowitz L (ed) Advances in experimental social psychology, vol 10. Academic Press, New York. https://doi.org/10.1016/S0065-2601(08)60358-5

Shaftel, H. (2019). Overview: Weather, Global Warming and Climate Change. Retrieved March 16, 2020, from https://climate.nasa.gov/resources/global-warming-vs-climate-change/

State of Youth. (2019). For Youth by Youth. Retrieved April 14, 2020, from https://stateofyouth.org/about

Steg, L., Bolderdijk, J. W., Keizer, K., & Perlaviciute, G. (2014). An Integrated Framework for Encouraging Pro-environmental Behaviour: The role of values, situational factors and goals. Journal of Environmental Psychology, 38, 104–115. https://doi.org/10.1016/j.jenvp.2014.01.002

Stern P.C., Dietz T, Abel T.D., Guagnano G.A., Kalof L. (1999) A value-belief-norm theory of support for social movements: the case of environmentalism. Human Ecology Review 6(2):81–97

Sustainable Development Solutions Network in collaboration with BertelsmannStiftung. (2019). Sustainable Development Report 2019. Retrieved from https://www.sdgindex.org/

Tharwat, A. (2018). Classification assessment methods. Applied Computing and Informatics, 1–13. https://doi.org/10.1016/j.aci.2018.08.003

The Royal Society. (2017). 6. Climate is always changing. Why is climate change of concern now? | Royal Society. Retrieved March 16, 2020, from https://royalsociety.org/topics-policy/projects/climate-change-evidence-causes/question-6/

The Royal Society. (2020). Climate Change: Evidence & Causes. Retrieved from https://royalsociety.org/-/media/Royal_Society_Content/policy/projects/climate-evidence-causes/climate-change-evidence-causes.pdf

Trope, Y., Liberman, N., 2010. Construal-level theory of psychological distance. Psychol. Rev. 117 (2), 440–4463. https://doi.org/10.1037/a0018963

U.N. (2015). The Millennium Development Goals Report 2015. Retrieved from https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf

UNFCC. (2018). What is the Paris Agreement? Retrieved March 10, 2020, from https://unfccc.int/process-and-meetings/the-paris-agreement/what-is-the-paris-agreement

Van Loon, A. F., Stahl, K., Di Baldassarre, G., Clark, J., Rangecroft, S., Wanders, N., ... Van Lanen, H. A. J. (2016). Drought in a human-modified world: reframing drought definitions,understanding, and analysis approaches. Hydrology and Earth System Sciences, 20(9), 3631–3650. https://doi.org/10.5194/hess-20-3631-2016

Watkins K. (2006) Human Development Report 2006-beyond scarcity: Power, poverty and the global water crisis. UNDP Human Development Report (Palgrave Macmillan, New York). Retrieved from https://www.undp.org/content/undp/en/home/librarypage/hdr/human-development-report-2006.html

White, K., Habib, R., & Hardisty, D. J. (2019). How to SHIFT Consumer Behaviors to be More Sustainable: A Literature Review and Guiding Framework. Journal of Marketing, 83(3), 22–49. https://doi.org/10.1177/0022242919825649

World Economic Forum. (2015). Global Risks 2015. 10th edition. Retrieved from http://www3.weforum.org/docs/WEF_Global_Risks_2015_Report15.pdf

Yong, A. G., & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. Tutorials in Quantitative Methods for Psychology, 9(2), 79–94. https://doi.org/10.20982/tqmp.09.2.p079

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Appendix

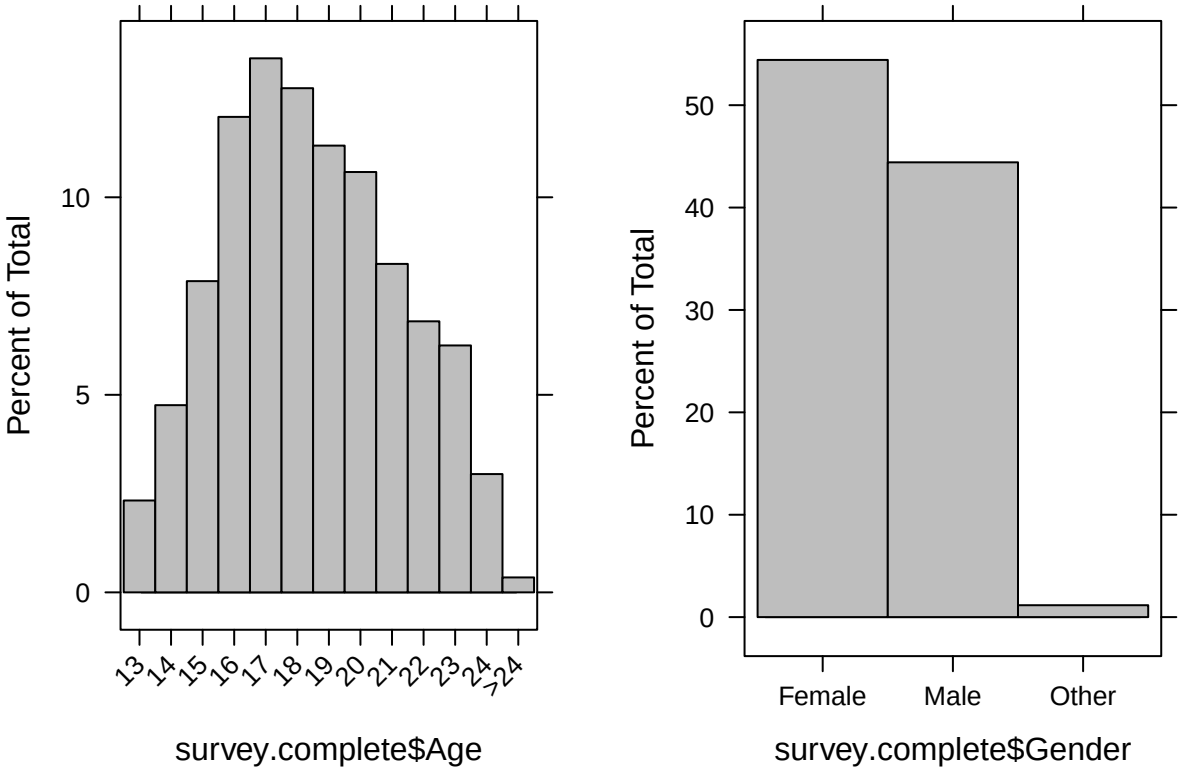

Figure 8: Histograms of variables Age (left) & Gender (right)

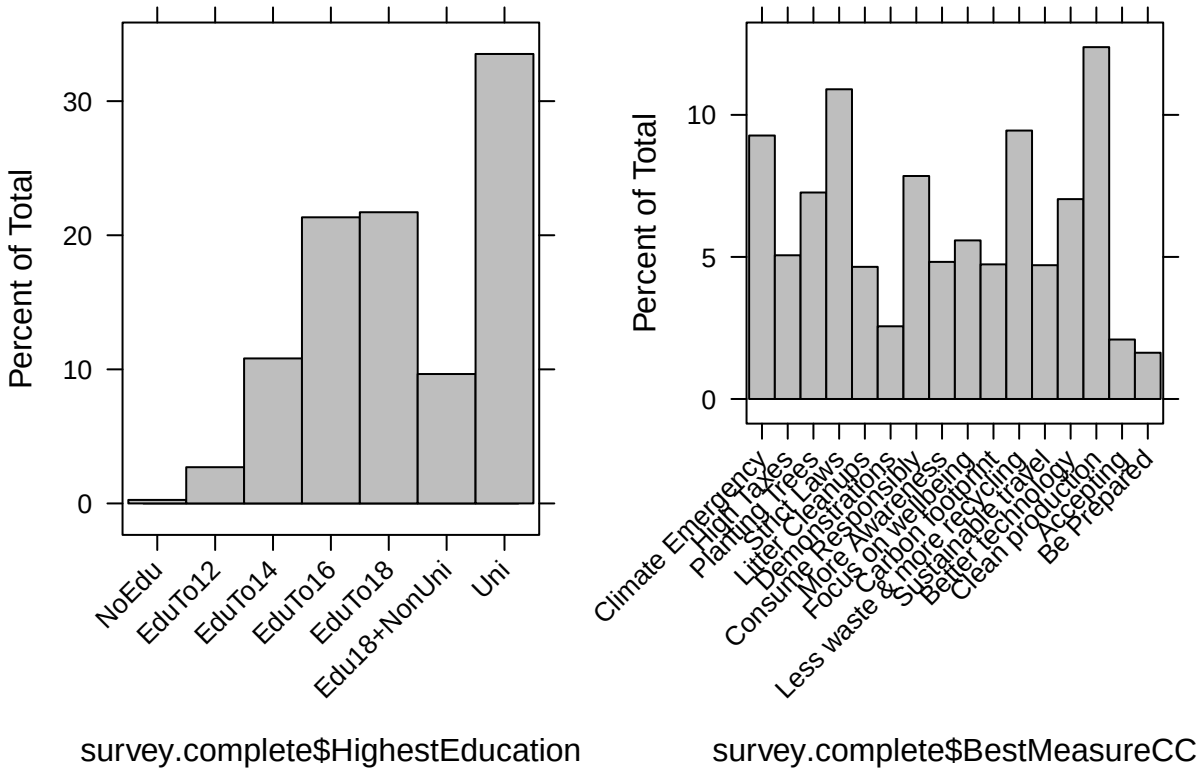

Figure 9: Histograms of variables HighestEducation (left) & BestMeasureCC (right)

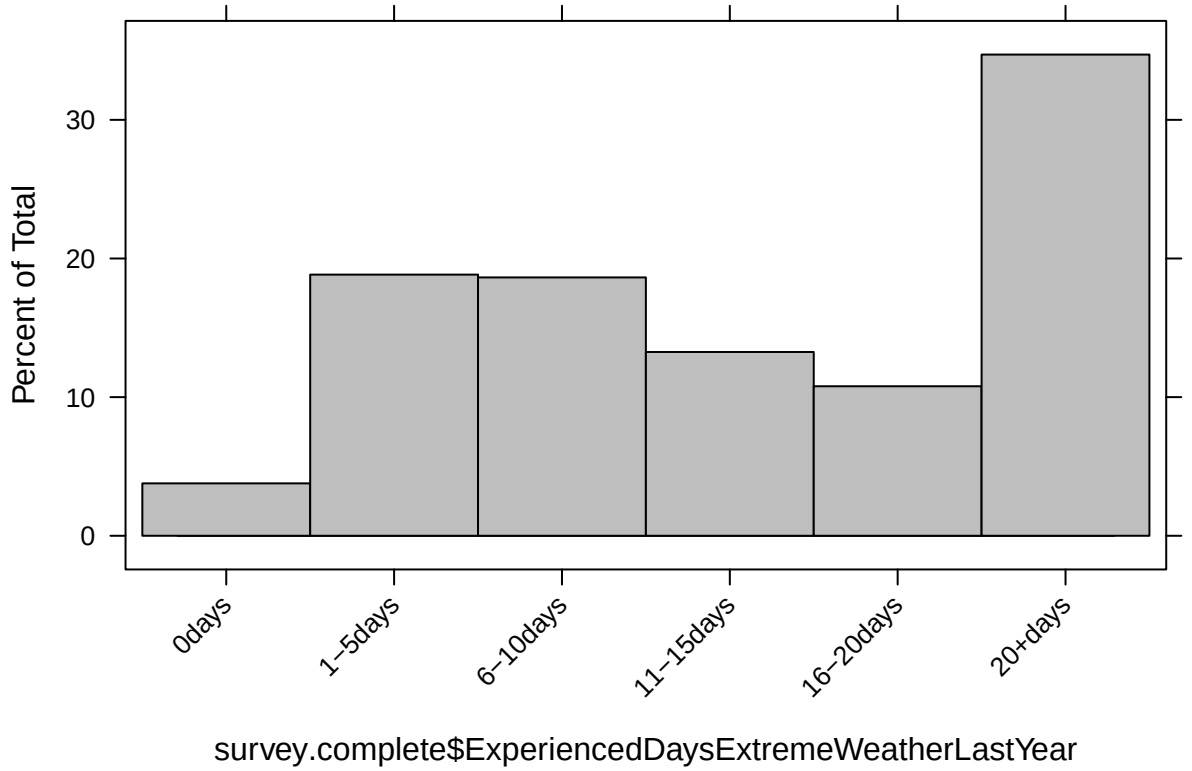

Figure 10: Histogram of variable ExperiencedDaysExtremeWeatherLastYear

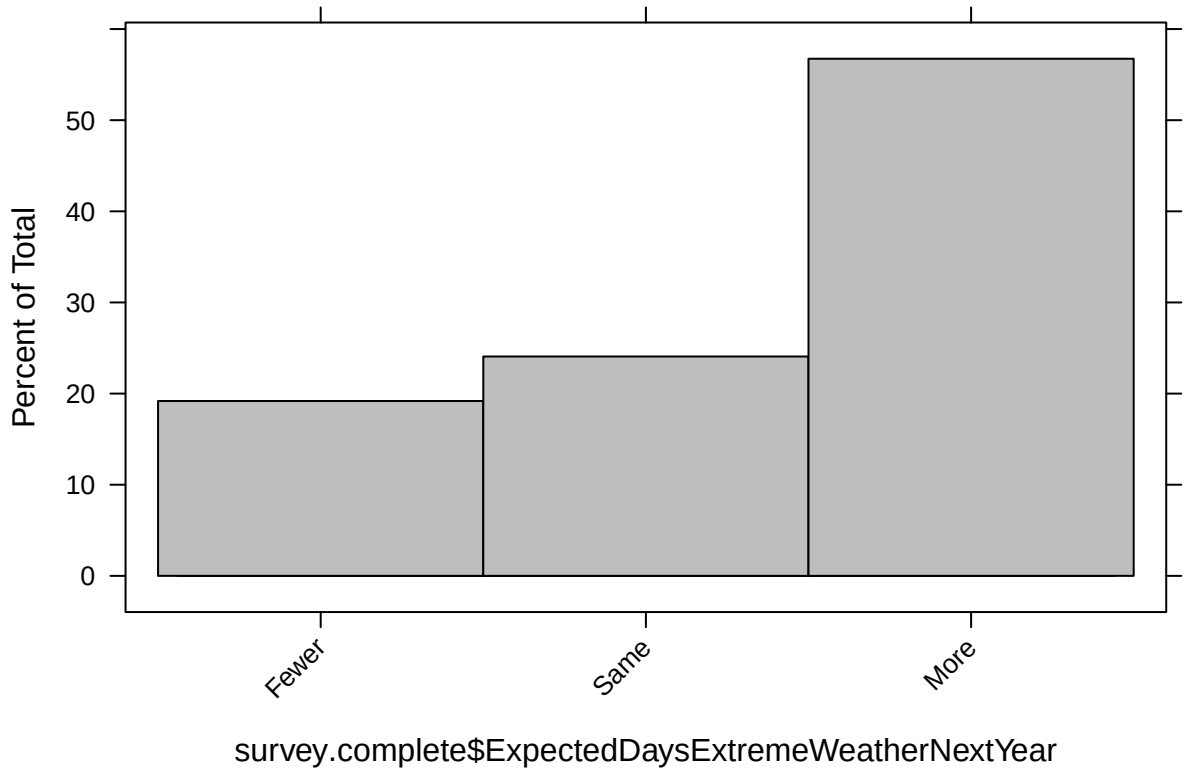

Figure 11: Histogram of variable ExpectedDaysExtremeWeatherNextYear

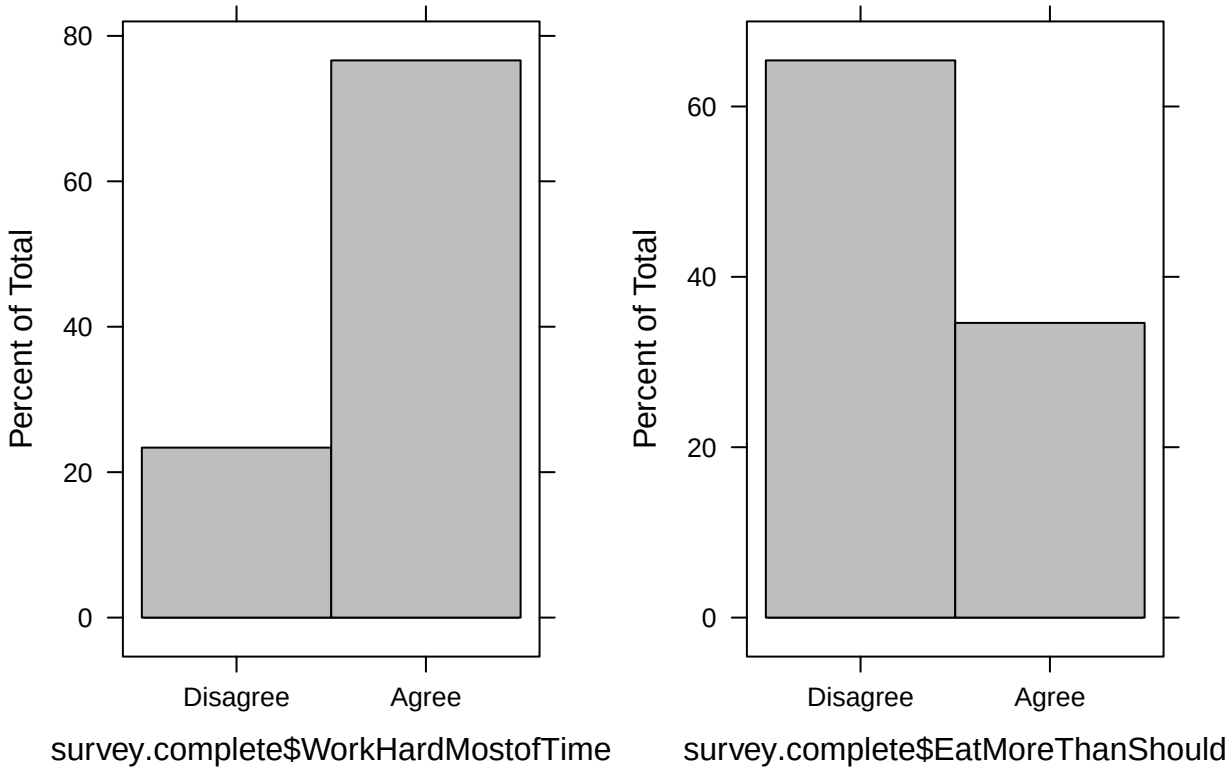

Figure 12: Histograms of variables WorkHardMostofTime (left) and EatMoreThanShould (right)

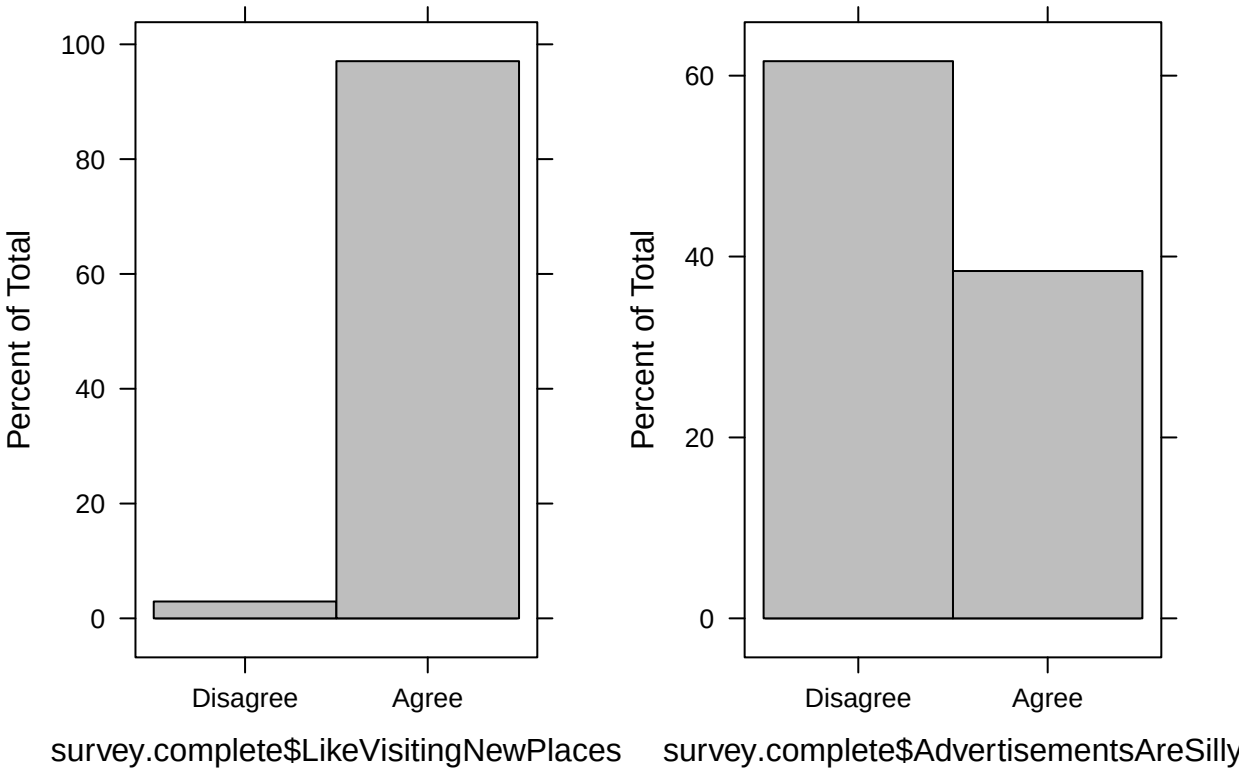

Figure 13: Histograms of variables LikeVisitingNewPlaces (left) and AdvertisementsAreSilly (right)

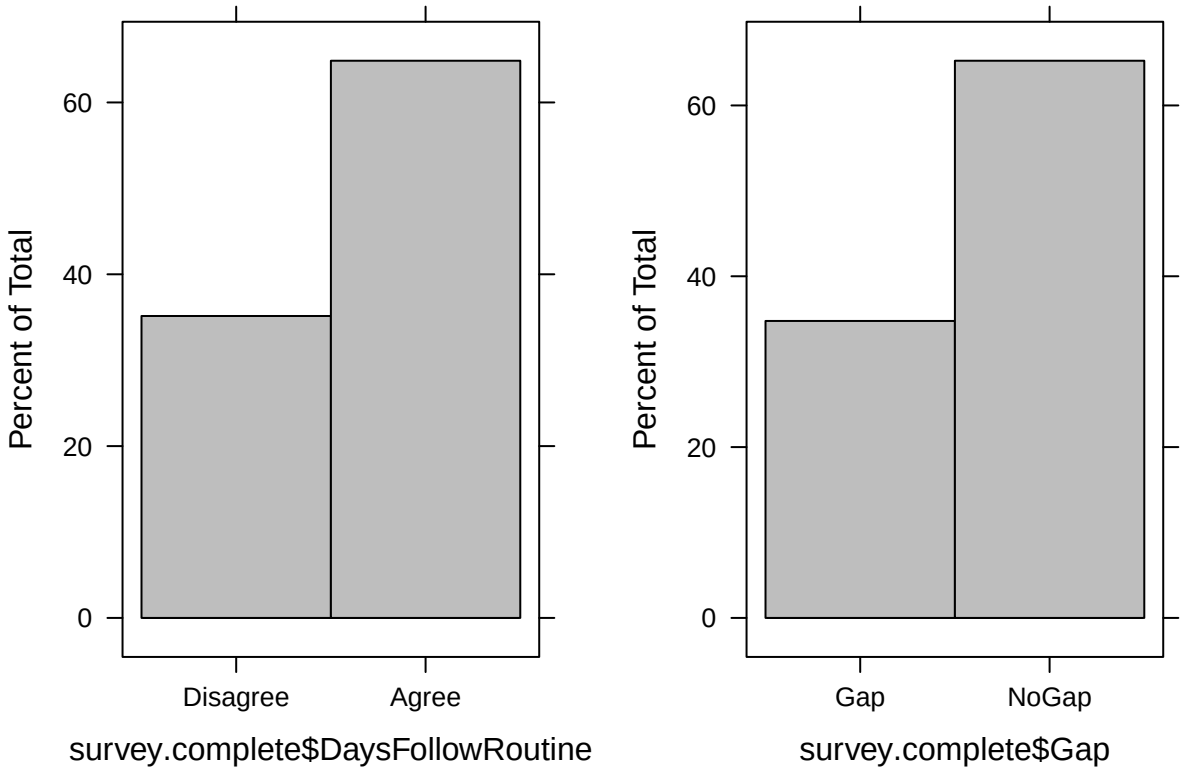

Figure 14: Histograms of variables DaysFollowRoutine (left) and Gap (right)

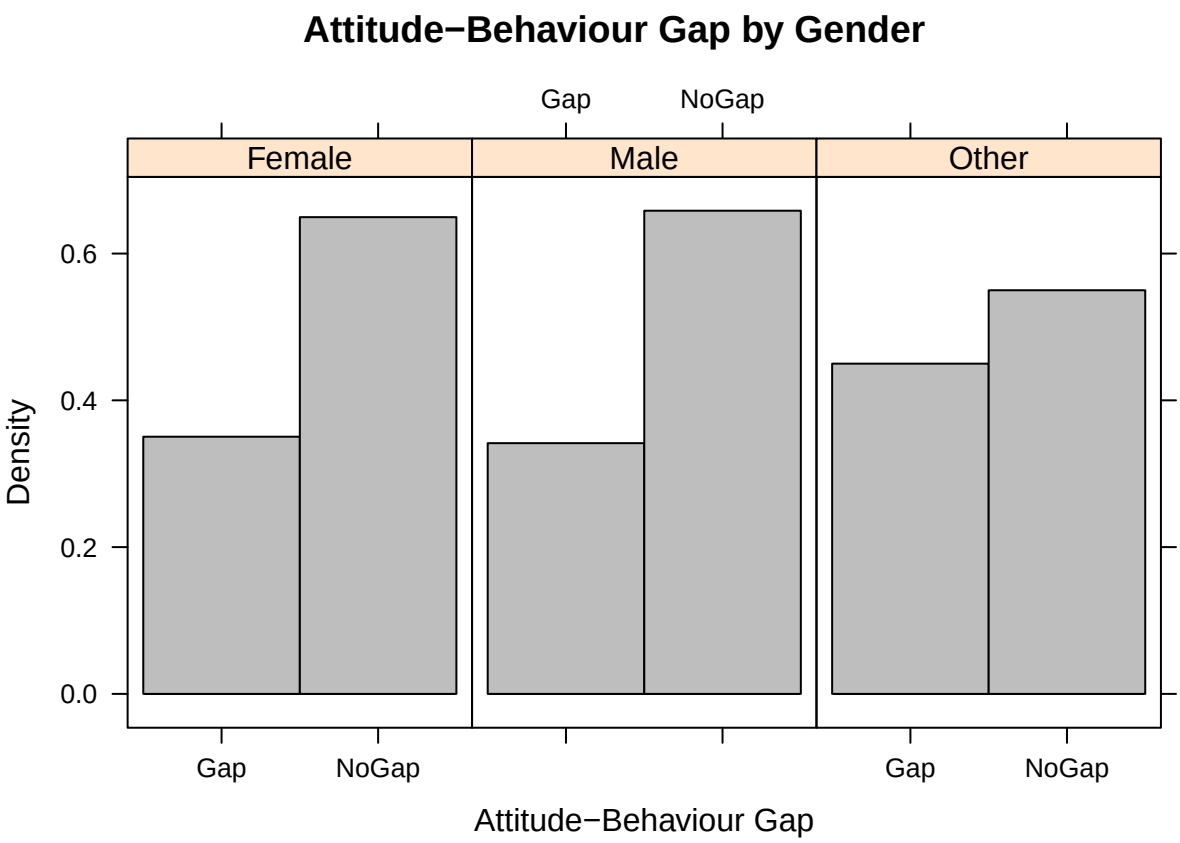

Figure 15: Histogram of Gap by Gender

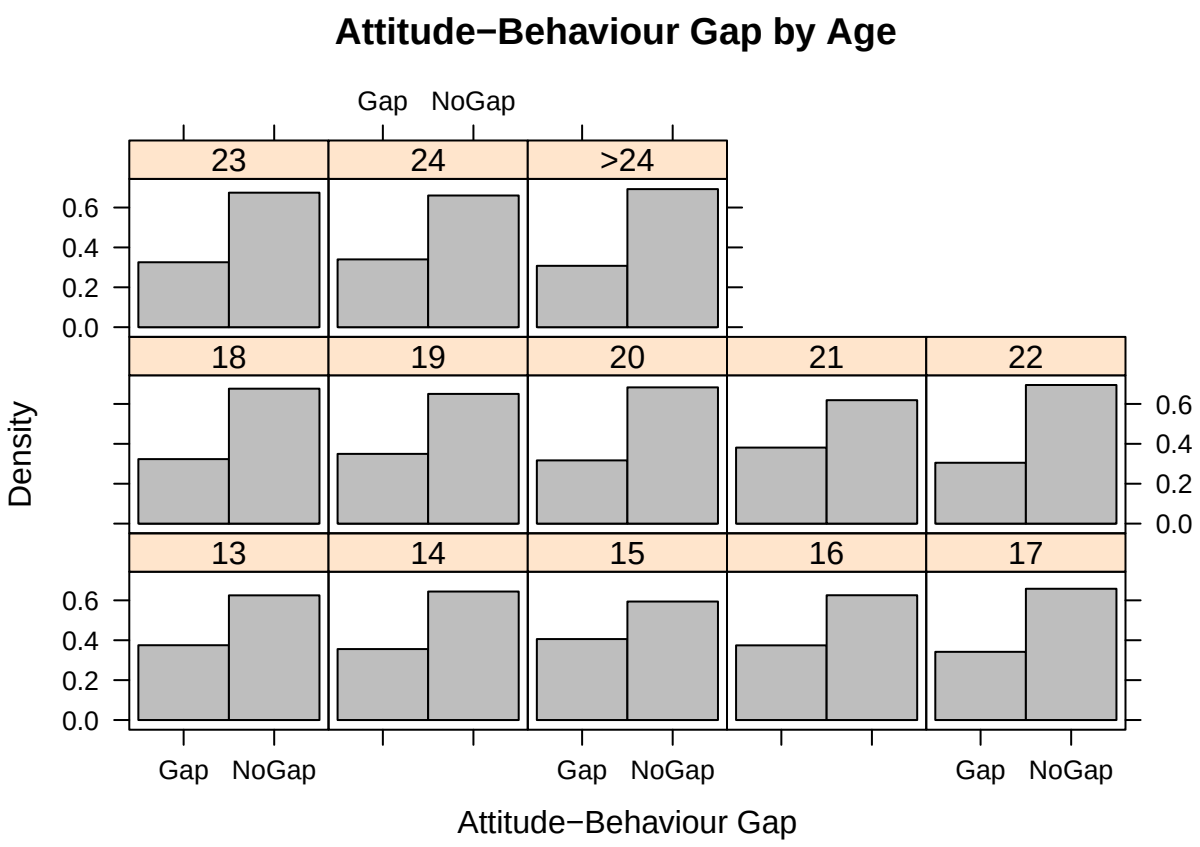

Figure 16: Histogram of Gap by Age

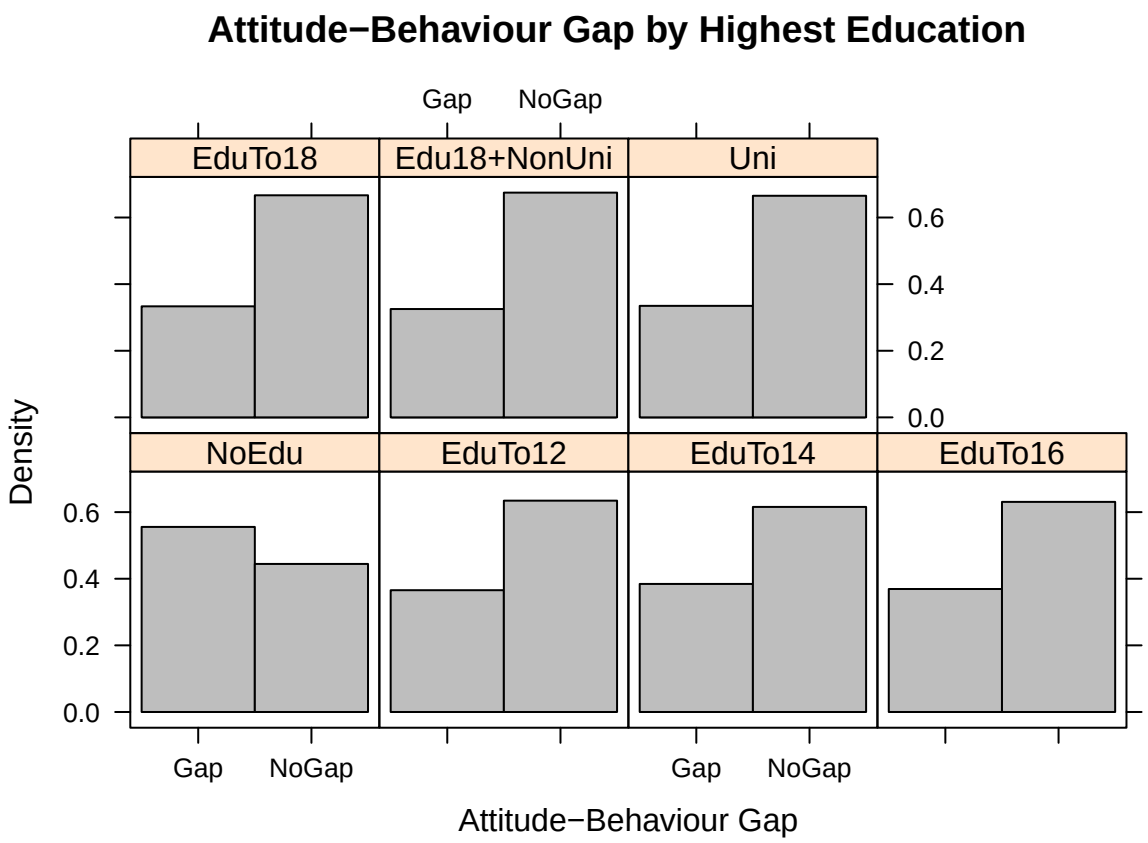

Figure 17: Histogram of Gap by Education