

Big Data Psychology

Rosa Lavelle-Hill

School of Psychology, University of Nottingham

N\LAB, University of Nottingham

Supervisors

James Goulding, N\LAB, University of Nottingham

Pete Bibby, School of Psychology, University of Nottingham

David Clarke, School of Psychology, University of Nottingham

Submitted Dec 2019

Abstract

This thesis investigates what big data can add to the psychological study of human behaviour; and how Psychological theory can inform developments in machine learning models predicting human behaviour. It works through the difficulties that arise when the fields of machine learning and psychology meet. While machine learning models deal well with big datasets, they are designed for prediction, neglecting psychologists' desire to, not just predict, but understand behaviour. Psychology does well at using theory to specify models and explain the variance within a sample, yet can fail to consider how transferable the findings are to new samples. This research harnesses over a million loyalty card transaction records from a high-street health and beauty retailer linked to 12,968 questionnaire responses measuring demographics, shopping motivations, and individual differences. Equipped with real world behavioural records, and information on potential psychological and demographic drivers of behaviour, this thesis explores the ways in which psychological research can be undergone using big data to better understand three main areas: well-being, environmental behaviours, and anxiety symptoms. This thesis has the goal of marrying the strengths of traditional psychological methodology (utilising theoretical knowledge, quantifying uncertainty, and building interpretable models) with the exciting possibilities afforded by big data, all whilst ensuring that the models are generalisable and do not overfit. The following chapters discuss and evaluate novel research in this space, as well as the difficulties encountered, and compromises made, in undertaking 'Big Data Psychology'.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors James Goulding, Pete Bibby, and David Clarke for their ongoing guidance, words of wisdom, and hilarious joint supervisor meetings.

James's energy spent helping me write papers, time debating scientific and statistical approaches in the pub, and his confidence in my abilities has made me a better writer, more rounded researcher, and allowed me to accomplish things I did not think were possible five years ago. I am grateful for all the opportunities and friendship he has given me, not to mention the family for which N\LAB has become over the years.

Pete's continuous enthusiasm for psychological research greatly inspired me during my undergraduate degree as my lecturer in Statistics and Conceptual and Historical Issues in Psychology. Throughout my PhD he has continued to be an incredible teacher and a fountain of knowledge. Am particularly thankful for his decision to take me on as a PhD student five years ago, and his willingness to offer supervisory support even after retirement. I will continue to be inspired by this strength of character, and strong opinions.

David's lectures in my undergraduate module Patterns of Action ignited my interest in studying human behaviour using different approaches. He has continued, and will always continue, to be an incredible inspiration to me as a researcher, and as a person. His passion for science and psychology, not to mention his warm nature (which first struck me during his university open day talk in 2011) has bought me great comfort and confidence over the years. I hope his mind will continue to inspire many students and people to come.

Besides my advisers, I would like to thank the N\LAB family for their jokes, games, and Aspire lunches. I am grateful to Gavin Smith for his time helping me with technical issues, and passionate discussions about variable importance in the pub. I am thankful to Bertrand for his willingness to help (and talk sport over lunch). I am also greatly appreciative of Katie, Vanja, Gregor, and Georgie for helping me stay positive, and bringing me laughter throughout this journey. Finally, I would particularly like to thank Maddy for being such a caring friend, and an immense source of joy every day.

Last but not least, I would like to thank my family. My parents, for offering their undying love and support, and encouraging me from a young age to pursue whatever makes me happy. My sister, for her inspiring strength and kindness. My grandparents, for their love, time, and support. And finally, Sam, for his incredible patience, stimulating conversations, love and companionship - you have made even the most stressful of times an absolute joy.

Contents

1	Introduction	9
1.1	Preview	9
1.2	What exactly is Big Data?	13
1.3	Big data as an opportunity for Psychology	15
1.3.1	Data at Scale	15
1.3.2	Naturalistic Behaviour Capture	20
1.3.3	Studying the Underrepresented	23
1.3.4	Preventing Overfitting	26
1.4	Prediction versus Explanation	26
1.5	Induction versus Deduction	28
1.6	Potential Barriers to Big Data Psychology	29
1.7	Overview of the Thesis	31
2	Background and Literature Review	34
2.1	What is Machine learning?	34
2.1.1	Why use a Machine Learning Approach?	35
2.1.2	Decision Trees	36
2.1.3	Random Forests	41
2.1.4	Cross Validation	42
2.2	What would Big Data Psychology look like?	43
2.3	A Review of Big Data Psychology Research	45
2.3.1	Data as a Behavioural Measure	46
2.3.2	Related/ Peripheral Fields’ use of Big Data	47
2.3.3	Marketing Sciences and Consumer Psychology	48
2.3.4	Literature Informing Psychological Theory	49
2.3.5	Big Data Methods and Tutorials	49
2.3.6	Summary of the Gaps in The Literature	51
3	Methodology and Data	53
3.1	A Review of the Factors Affecting Consumer Behaviour	54
3.1.1	Demographic Features	55
3.1.2	Personality Theory	57
3.1.3	Impulsivity	61

3.1.4	Self-Control	62
3.1.5	Time Preferences	62
3.1.6	Shopping Impulsivity	64
3.1.7	Frugality	66
3.1.8	Brand Personality and Brand Commitment	67
3.1.9	Summary	68
3.2	The Data	68
3.2.1	Loyalty Card Transaction Data	68
3.2.2	Survey Data	69
3.2.3	Survey Data Descriptives	70
3.2.4	The Survey Sample's Transaction Data	78
4	Personality and Consumer Behaviour Predicts Subjective Well-being	79
4.1	Introduction	79
4.1.1	Money and Happiness	79
4.2	Matz et al. (2016), a Replication: Buying things that match your personality makes you happier	81
4.2.1	Introduction	81
4.2.2	Method	82
4.2.3	Results	85
4.2.4	Discussion	87
4.2.5	Summary	90
4.3	Buying what people like you like: Consumer choice, personality, and well-being in big data	90
4.3.1	Introduction	91
4.3.2	Method	94
4.3.3	Results	97
4.3.4	Discussion	106
5	Predicting Well-being in Big Data using Machine Learning Methods	112
5.1	Introduction	112
5.2	Using Machine Learning to Gain Additional Insights	114
5.2.1	Insight 1: Fitting Models can lead to Inflated Effect Sizes	114
5.2.2	Insight 2: Decision Tree's can be used to Gain Additional Understanding	118
5.3	Summary	121
6	Understanding the Predictors of Plastic Bag Consumption in Big Data	123
6.1	Introduction	124
6.1.1	Plastic Bag Charges	124
6.1.2	Algorithm Modelling versus Data Modelling	128

6.2	Psychological and Demographic Predictors of Plastic Bag Consumption in Big Data	131
6.2.1	Aim	131
6.2.2	Predictions	132
6.3	Analysis 1	133
6.3.1	Method	133
6.3.2	Results	137
6.3.3	Discussion	142
6.4	Analysis 2	148
6.4.1	Method	149
6.4.2	Results	152
6.4.3	Discussion	155
6.5	Joint Discussion	158
6.5.1	Methodological Discussion	158
6.5.2	Limitations	162
6.5.3	Implications and Summary	162
6.6	Conclusion	163
7	Alternative Methods to Understand Feature Importance for Plastic Bag Consumption	165
7.1	Feature Importance Insights from Decision Trees and Random Forests	165
7.1.1	Decision Tree Analysis	166
7.1.2	Insights from the Tree	168
7.2	Feature Importance insights from a Random Forest	169
7.3	Feature Importance Insights from NMF	171
7.3.1	NMF Analysis	171
7.4	Discussion	174
7.4.1	Is Statistical Significance Important?	175
7.5	Conclusion	175
8	Investigating a Behavioural Measure of Anxiety in Transaction Data	180
8.1	Introduction	181
8.1.1	Method	185
8.1.2	Results	187
8.1.3	Discussion	192
8.2	Geospatial Analysis in Big Data	197
8.3	Do Boots Customers' Personality Traits Cluster Spatially?	198
8.3.1	Introduction	199
8.3.2	Analytical method	200
8.3.3	Results	202
8.3.4	Discussion	209

8.4	Data Comparison's	209
8.4.1	London Personality Comparison's to the BBC Personality Data	210
8.4.2	London Demographic Comparisons to the Census Data	212
8.4.3	Discussion	214
8.5	Investigating the Geographic Relationship Between Psychological Traits and Anxiety Medication Sales	215
8.5.1	Results	216
8.5.2	Discussion	218
9	Discussion and Conclusions	219
9.1	Advantages of using Big Data in Psychological Research	219
9.1.1	High Statistical Power	220
9.1.2	Big Data Enables the Use of Machine Learning Algorithms	220
9.1.3	Sampling Variability	221
9.1.4	Naturalistic Behaviour Capture	222
9.2	Limitations of Big Data in Psychology	223
9.2.1	There Can Still be Sampling Bias in Big Data	224
9.2.2	Machine Learning Algorithms can Discriminate Against Sub-populations	225
9.2.3	Traditional Statistics for Quantifying "Uncertainty" are Not Useful on Big Data	227
9.2.4	Visualising and Checking Data is Harder with Big Data	229
9.2.5	Summary	230
9.3	A Discussion on Effect Sizes in Big Data Psychology	231
9.4	A Discussion on the Limitations of using Survey Data	233
9.5	Ethics and Big Data	234
9.5.1	Ethics in a 'Big Data Society'	234
9.5.2	Ethics and Big Data Research	235
9.5.3	Big Data Psychology Ethics	236
9.6	Conclusions on the Scientific Approach to Big Data Psychology	238
9.6.1	Prediction versus Explanation	238
9.6.2	Induction versus Deduction	244
9.7	Concluding Remarks	246
10	Supplementary Materials	286
10.1	Chapter 3: Methodology and Data	286
10.1.1	The Personality Questionnaire in Full	286
10.2	Chapter 4: Personality and Consumer Behaviour Predicts Subjective Well-being	286
10.2.1	Preliminary Analysis: Using Personality Alignment to Predict the Amount Spent on Brands and Products	286
10.2.2	Tables of Brand and Product Descriptives	289

10.2.3	Personality Distributions	293
10.2.4	Expected Brand Personality	294
10.2.5	Personality of Participants in the Sample for Analysis 2 (Products) . .	295
10.2.6	Expected Product Personality	296
10.2.7	Correlations	296
10.3	Chapter 6: Understanding the Predictors of Plastic Bag Consumption in Big Data, Analysis 1	299
10.3.1	Analysis 1: Meta-Parameters	299
10.3.2	Analysis 1: A list of the Non-Significant Predictor Variables in the Logistic Regression Model	299
10.3.3	Analysis 1: Structural Equation Model Pathway Analysis	302
10.4	Chapter 6: Understanding the Predictors of Plastic Bag Consumption in Big Data, Analysis 2	303
10.4.1	Analysis 2: Meta-Parameters	303
10.4.2	Analysis 2: Full Results Table from the Logistic Regression Analysis .	303
10.5	Chapter 8: Investigating a Behavioural Measure of Anxiety in Transaction Data	306
10.5.1	Full List of Over the Counter Anti-Anxiety Medication used in the Study	306
10.5.2	A Full Table of Results from the Logistic Regression Model	307

List of Abbreviations

Abbreviation	Definition
IV	Independent Variable
DV	Dependent Variable
LRM	Linear Regression Model
MLR	Multiple Linear Regression
DT	Decision Tree
RF	Random Forest
NN	Neural Network
SEM	Structural Equation Model
CV	Cross Validation
VIF	Variable Inflation Factor
SWB	Subjective Well-Being
O	Openness to New Experiences
C	Conscientiousness
E	Extraversion
A	Agreeableness
N	Neuroticism
ES	Emotional Stability
TIPI	Ten Item Personality Inventory
BIS	Behavioural Inhibition System
BAS	Behavioural Activation System (a measure of impulsivity)
BAS-RR	Reward Responsiveness
BAS-FS	Fun Seeking
BAS-D	Drive
SM	Shopping Motivation
b-p (match)	basket-participant (match)
EPP	Expected Personality of Purchase
PA	Personality Alignment
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
FDR	False Discovery Rate

Chapter 1

Introduction

1.1 Preview

New technologies and digital services are increasingly allowing humans to create rich digital footprints. The use of wearable technologies, digital records, and social media means we leave an incredible amount of data in our wake. This “datafication” of modern life, combined with modern computing power, and easily transmissible statistical knowledge, creates opportunities for new ways to study human behaviour. Using mass datasets, we are able to detect and study subgroups within society that were not previously analysable in smaller samples. Powerful pattern-detection algorithms can identify trends that humans alone, without the aid computers, can not. The combination of big data and algorithmic modelling has the potential to provide us with new insights on human behaviour, giving us the opportunity to advance our understanding of human nature.

Psychology, the science of understanding the brain and behaviour, traditionally deals with small datasets. Behaviour is commonly measured on an individual level and in a controlled artificial environment. Behaviour, in all its complexity, is thus reduced to small, mirco-level tasks, reaction times, or decisions. These can be statistically related to the size and structure of brain regions, to individual cells firing in the brain, or to the flow of oxygen, sugar glucose, and electricity over specific cortical regions. The brain controls behaviour, and many in psychology take the stance that to understand behaviour we need to focus on its origin. But just as the 1s and 0s in a computer program are incomprehensible to a human, the cells firing in a brain can not describe human behaviour as we experience it. Humans are the most sociable animals on the planet, and to study them in isolation is surely limiting our ability to understand human behaviour. When we think about our memories and experiences that are most important in life, what makes us most human - love, parenthood, culture, sex, adventure - these are a long way removed from reaction times, forced choices, and memory tasks used to measure behaviour in the lab.

Controls are required in studies of human behaviour to assign causality to the independent variable being tested, reducing the likelihood of an effect being attributed to some unmeasured external factor. This manner of studying individuals is time and resource expensive, motivating the collection of sample sizes just large enough to find the predicted

effect size, and no bigger. In addition, the samples collected in psychological research mostly comprise of just a small slice of humanity - whomever university researchers can get their hands on.

In the era of big data, human behaviour can be quantified on a large scale, as it exists in the real world. The large sample sizes allow for the modelling of a greater complexity of interactions, and the uncovering of more subtle effects. The release of big datasets on human behaviour should be a treasure trove of opportunities for psychologists looking to study naturalistic human behaviour on a large scale. Not only does big data negate the need for additional data collection by researchers, it quantifies behaviours unobtrusively, reducing social desirability biases that occur when a person is explicitly aware they are being monitored. Despite this, the field of psychology has been slower on the uptake of big data research than other disciplines like computer science, engineering, or marketing. Such disciplines are used to working with large datasets, and thus are more readily equipped with the skills, models, and statistics for big data research. For psychologists, exploiting big data requires rewriting the research methods and statistics textbooks that most in the discipline were raised on. Even greater, it requires a monumental shift in the conceptualisation of how behaviour research should be conducted.

A deductive approach to science, as per Popper (1949), is commonplace in psychology, born from an era, pre-datafication, when data was hard to get hold of. Deduction entails that researchers formulate a research hypothesis they want to test, and set out to collect data. When data collection is resource expensive, this produces datasets with a small number of participants, often from one geographic area and from a narrow window in time, datasets with only a few variables measured and used to test a small number of hypotheses. Such data lends itself nicely to inferential statistics (Faraway and Augustin, 2018). With big data research, however, the data is already there, often with many measured variables. This increases the potential for knowledge discovery, but also the likelihood that variables are related to one another. Datasets with high multicollinearity are not conducive to traditional statistical models commonplace in psychology, such as linear regression, where the interpretation of β coefficients becomes spurious (Minitab, 2013).

When research is being conducted in an environment where data is resource expensive, conditions don't lend themselves well to keeping spare data aside, or reducing the data available for modelling. Yet this is something that is required to examine the generalisability of the models results. Thus, in psychology, models are fitted to a sample to explain the variance within that sample, rather than to predict new (held-out) values. In big data, the larger the sample size, and the greater the complexity of the model (i.e. the more variables), the greater the likelihood of the model 'overfitting' to a sample - modelling the noise, as well as the signal, within the data. Therefore, in big data research using models to predict out-of-sample data, a method rarely used in psychology, becomes vital in big data research.

More than the actual size of big data, a defining feature is its potential to use with machine learning and other advanced data analytic methods to extract patterns, detect signals,

and address questions that are difficult to address with smaller datasets. In neuroscience, computational models are being exploited to model the brain’s processes, such as memory (Blakeman and Mareschal, 2020), to help researchers to shed light on how such complex systems are operating. Psychology can look to do the same with behaviour - by modelling human behaviour with the aid of complex models, we can start to identify and unravel relationships which were not previously visible using statistical analysis on small data.

Behaviour is the production of complex systems of interacting entities. Perhaps the root of behaviour is in the brain, but which parts of the brain, and how do they interact with other parts, the environment, past experiences, other people, other systems, and order and randomness in the world? Human behaviour does not occur in isolation, and the world is not controlled. Real world behaviour is a complex, chaotic, and evolves temporally. Larger datasets, combined with the use of more complex models and a greater focus on generalizability, can only help to model this situation. And like pixels in an image, the more data points we can obtain and supply our models with, the clearer the picture we ought be able to obtain.

This Thesis

This thesis investigates what big data can add to the psychological study of human behaviour; and how Psychological theory can inform developments in machine learning models predicting human behaviour. It works through the difficulties that arise when the fields of machine learning and psychology meet. While machine learning models deal well with big datasets, they are designed for prediction, neglecting psychologists’ desire to, not just predict, but understand behaviour. Psychology, can do well at accounting for the variance within a sample, yet fail to consider how transferable the findings are to new samples. On the other hand, when theory and domain knowledge is disregarded completely, a model can produce an accurate prediction with a theoretically implausible explanation. The following chapters discuss the difficulties encountered, and compromises made, in undertaking psychological research using a large transaction database. This thesis has the goal of marrying the strengths of traditional psychological methodology (utilising theoretical knowledge, quantifying uncertainty, and building interpretable models) with the exciting possibilities afforded by big data. At the same time it explores the need to ensure that models are generalisable and do not overfit.

This research specifically exploits a dataset of loyalty card users’ transaction data. Loyalty cards offer consumers benefits or rewards in exchange for the retailer being able to track the individual’s spending. This data is typically used to inform marketing and advertising strategies. A loyalty card ties all the person’s purchases (for which they use the card) to one ID - allowing for their behaviour to be tracked over time and across stores. The data used in this research is from Wallgreen Boots Alliance (from here on in simply referred to as “Boots”), a multi-national health and beauty retailer, selling everything from food and health products to make-up, toiletries, and nappies. The brands, products, and quantities

someone buys can be an indicator of a multitude of different health and environmental behaviours. For example, if someone buys smoking cessation products we can infer from this one transaction that they do/used to smoke and are trying/have quit. From the frequency and pattern of their transactions we can gain an understanding of whether they were able to quit smoking or not, how long it took them, and how many attempts. From such analyses, inferences about the type of person they are can be made. In this way, a person's shopping can carry indicators of their psychological states and traits - including Subjective Well-Being (SWB), impulsiveness, and extraversion (sociability) - and this is an area this thesis explores.

Using the transaction data, for each individual, a number of different behaviours can now potentially be examined as they happen in the real world. To better understand the drivers of an individual's behaviour we require data, not just on what they have done, but on the possible drivers of their behaviour. In this thesis, mass survey responses were collected from loyalty card users which could be linked to their behavioural transaction data, and hence explore such drivers. From a questionnaire, information was gained on a subset of the loyalty card users' demographics and personality traits. Equipped with real world behavioural records from the transaction data, and information on the potential psychological and demographic drivers of behaviour from the personality questionnaire, this thesis sets out to explore the ways in which psychological research can be undergone in big data.

Chapter 1 of the thesis introduces the nature and idiosyncrasies of big data, and the ways in which psychology can look to harness it. Chapter 2 introduces machine learning and the methods used in this research, namely cross validation, decision trees, and random forests. It then describes the space where psychology and machine learning meet, referred to as *big data psychology*, and reviews the previous literature conducted in this sphere. Chapter 3 introduces the survey distributed to customers to collect information on consumer's psychological and demographic characteristics, and provides an overview of the literature used to construct the survey, covering the different psychological and demographic predictors which can influence consumer behaviour. It then introduces the data - both the transaction data, the survey data, and the sample when both are linked together. The data is described and descriptive statistics are reported. Chapter 4, the first of five analytical chapters, uses the brands and products people buy to predict subjective well-being using linear regressions. Chapter 5 then explores how machine learning methods can potentially enhance the investigation of well-being in transaction data. Chapter 6 examines methods to combine machine learning and statistics to uncover the psychological and demographic predictors of single-use plastic bag buying. And Chapter 7 explores how decision trees, random forests, and non-negative matrix factorisation can aid variable importance when multicollinearity is present in the data. In Chapter 8, the thesis explores whether the transaction data can be used to create a behavioural measure of anxiety, and how big data can be linked to other open datasets to study personality and behaviour geographically. Finally, Chapter 9 discusses the advantages and limitations of conducting psychological research in big data, and how

psychology might move forward as a ‘big data discipline’.

1.2 What exactly is Big Data?

Big data involves the retrieval, storing, and analysis of large amounts of digital information. Big data research or analysis can also be referred to as data mining, knowledge discovery in databases, data or predictive analytics, or generally as data science. Since the 1990s the interest in big data has been well documented in the academic literature (Harlow and Oswald, 2016). Big data was traditionally associated with computer science, statistics, marketing, and business, but over the last five years big data has been increasingly used by researchers in social or the behavioural sciences.

This shift has only occurred recently with the introduction of cloud computing and distributed computing providing the computing power to store and analyse such large quantities of data. 90% of all data has been created in the last two years (according to IBM (2017)). And over the course of 2019 the big data market was expected to grow by a further 17% (Statistica, 2018). While big data requires specific infrastructure and software to store and process such large amounts of information, recent years have seen the development of more open source software making big data analysis more accessible. (From MySQL and Postgres to Hadoop, an open-source framework for storing data and running code requiring large processing power, using distributed clusters of hardware.)

The defining feature of big data as an umbrella term is not only its size, but that its generation is automated - making it quick and easy to acquire. Although how ‘big’ data has to be before it is called ‘big data’ is not clearly defined. Big data is often described on 5 dimensions (Cuartas Arias, 2017; Feldman et al., 2012; Clifford, 2008): the volume or quantity of data; the speed the data can be acquired or scraped (i.e. from web pages, social media or digital applications); the variety of the data i.e. the diversity of the digital resources producing the data, and the format the data can be in - whether it’s unstructured (irregular data that doesn’t have a specific form, like that scraped from a website) or structured (in lists and tables); its veracity - how useful and trustworthy the data is; and finally its value (or the worth of the extractable insights/information) to various stakeholders/ decision makers. However these definitions originate from marketing pitches and the impact big data has on research, from psychology to behavioural science, is less well defined.

Typically big data requires some pre-processing steps in order to clean it, and get it into a form which can be modelled or analysed statistically. I will argue in this thesis that the quality that makes big data distinct from small data is not merely its size but its function, and the type of complex algorithmic analytic methods that can be used with it. Such methods can be so complex that humans cannot interpret exactly what they are doing - they are a ‘black box’. Yet their growth in popularity has led to the development of ‘plug in’ software packages such as SciKit-learn in python (Pedregosa et al., 2011) that now makes machine learning tools more accessible for everyone in research. Thus, more people are able

to address questions using big data, that were perhaps previously difficult to address with small datasets.

The big data analytics market is itself large, and estimated to reach \$103 billion by 2023 (Petrov, 2019). But just because data is there, doesn't mean it is there for *everyone*. Famously, data has been described as the “new oil of the internet” by Alex Pentland, Toshiba professor of Media Arts and Sciences at MIT (2014). Viewed as highly valuable, those who possess it in abundance can profit - “It’s a new asset class, a new value, a new money” (Petrov, 2014). Just as organisations keep company secrets close guarded, datasets are usually kept under lock and key. 97.2% of organizations are investing in big data and AI (NVP, 2018), but rich behavioural datasets are hard for researchers to get their hands on. Just as in business, research labs are in competition with one another over grant money and there is little incentive, other than the progression of science, to share such datasets. In most cases, even if researchers wanted to share data, privacy issues would prevent it. Some datasets exist online as open source and available to download. For example, the UK census data, and both the UK (Data.gov.uk) and the US government data (Data.gov) are online. Psychological datasets also exist: The Big Personality Test by the BBC (University of Cambridge, 2015) stores personality data of UK participants; the myPersonality.org database (Kosinski et al., 2015) stores personality as well as the Facebook profile data of over six million participants; and The Stanford Network Analysis Project web-site (Leskovec and Krevl, 2014) hosts a wide range of datasets, including social networks, Tweets, and product reviews. Such open datasets are anonymised or aggregated for data security, which means researchers cannot contact the people with further research questions or surveys to link to their data. All in all, although digital data is omnipresent, because of privacy issues and ‘owning’ data holds a massive competitive advantage, the use of individual level data for behavioural research is still in its prime.

In recent years the academic disciplines of genomics, medicine, and environmental sciences have been the biggest proponents of big data research (Harlow and Oswald, 2016). Environmental sciences are particularly concerned with visually representing data spatially using geographic information systems (GIS). On the back of growing interest in this field, in 2013 an open access and peer-reviewed journal *Big Data* was founded. The journal advertises a comprehensive coverage and audience, although mentions of psychology and the broader social sciences in general are limited. In 2014 two further journals were founded - *Journal of Big Data*, and *Big Data Research* - neither of which show any strong orientation towards the social sciences. Early proponents of big data in psychology, Harlow and Oswald (2016), don't blame the omission of psychology in the field of big data on the actors in the field, instead they affirm that psychology “should be proactive and take advantage of a real opportunity in front of them” (the reasons why psychology has been slow off the mark with big data research will be explored further in Section 1.6).

1.3 Big data as an opportunity for Psychology

Big data is an opportunity for psychologists (Harlow and Oswald, 2016) for a number of reasons. In this chapter I will consider potential advantages/disadvantages of conducting psychological research in big data under four overarching themes:

1) Data at scale. Big data can be acquired quickly, cheaply, and on a large scale. Mass datasets can be exploited to help solve issues such as low statistical power which have been attributed in part to the small sample sizes generally used in psychological research. Big data being cheap and easy to acquire *en mass* has the potential to alleviate the time and money resource costs of longitudinal and or multi-cultural research, which puts many funding bodies and researchers off conducting such large scale studies. Big data is often collected over multiple time points, and often has spatial information attached to it, allowing the analysis of behaviour and trends over time with minimal extra resources.

2) Naturalistic behaviour capture. Big data can be used to capture behaviour occurring outside of laboratory conditions so that results generalise better to the real world.

3) Studying the underrepresented. More fine grained data on people from more diverse backgrounds can be used to better understand minority groups in society who have been previously overlooked due to a lack of information. More data may help to reduce discrimination occurring from treating everyone the same as the strongly represented groups. For example, big data could help to reduce error from sampling bias by helping to reduce the reliance on WEIRD populations (Western, Educated, Industrialised, Rich and Democratic) (Henrich et al., 2010).

4) Preventing overfitting. Big data and machine learning methods can help researchers to tackle the replicability crisis that is occurring in psychology (Open Science Collaboration, 2015). By having enough data to test models out-of-sample, or to recycle data, researchers can ensure they are not modelling noise in the sample, and can ensure generalisability to a broader population.

The following section outlines each of these opportunities for psychology in more detail.

1.3.1 Data at Scale

The Problems with Small Samples Sizes

Samples are used in psychology to gain an estimate of an effect within a population. It is often impossible, or at the least extremely expensive to sample every person within a population, and ultimately not necessary. For example, time and money would be wasted sampling everyone in the UK to check with absolute certainty whether on average men are taller than females. Thus, samples are selected to *represent* a population, and statistical techniques exploited. However, herein lies many issues, not least sampling variability. When we draw different samples from the population, there is variability in the parameter estimates based on the number and type of people drawn. The more people in the sample, the more likely the sample estimates will match the population estimates (given the assumptions that

the distribution of the sample matches the distribution of the population, and there is no trade-off between data quality and quantity).

To interpret results (particularly confidence intervals and p -values), it is assumed that the data can exist on its own (i.e. is not collected for a meta-analysis). One can interpret the findings to exist only within the sample collected. For example, “Out of the 100 people surveyed in the doctors waiting room, the majority of them are under 20 years old”. But this is not generally interesting or useful, unless researchers were particularly interested in those specific individuals in the sample (i.e. they have some rare/unique condition that is interesting to study). The results can be extrapolated to a coarser level “In Nottingham more young people go to the doctors than older people”. The age of the people who go to the doctors in Nottingham, however, is a niche area of interest, and perhaps it is more interesting to study this effect across a ‘wider’ or the ‘general’ population, such as that of the UK. Further, leaps in generalisability, without taking into account the context can lead to false conclusions being made. For example, perhaps the GP surgery in the study is actually a university surgery, and it is freshers week, thus freshers flu is plentiful. Because of this, there are more young people than old people in the surgery on this day. Here, the temporal and geographic context can explain the finding. In this case, the model is correct given the data. But the sample is limited, and the model is incomplete (it has no contextual variables), leading to false conclusions being made.

A sample that is representative of the population doesn’t necessarily need to be the largest sample possible. People can be drawn into a sample along the population distribution of a few key variables (for example, ensuring 50% males and 50% females (the population distribution) exists in the sample). Ensuring a good representation of different religions, income groups, or males and females creates a more varied set of people to study. Recording information on gender, for example, allows the researcher to test whether the observed effects are the same for both males and females; or whether an extraneous variable can explain the effect. However, potentially confounding variables may be unmeasured. If the researcher is not aware of them, they can not be explicitly controlled for in the model. Yet these hidden or latent factors could be responsible for the effect on the dependent variable. For example, by chance there could be two dyslexic individuals in a sample of 20: if two of these end up in condition A ($N=10$) (rather than equally distributed over A and B) this could be responsible for on average sample A performing worse in a reading test. If Dyslexia was not measured and controlled for, incorrect conclusions can be made on the effect of the independent variable (condition) on the dependent variable (reading ability).

The larger the sample selected, the greater the chance of selecting a greater variety of individuals. Sample variability is a particular issue within psychology, as the effects researchers are looking for are generally small to moderate. In small samples, randomness or noise can be mistaken as small effects. To reduce the likelihood of this occurring, researchers can directly measure potentially confounding variables to control for them statistically in their model. However, measuring too many confounding variables in a small sample can also

produce spurious results. When the number of variables measured in the dataset is large, and the number of observations is small (characterising a ‘wide’ dataset), models such as regression can fail to produce sensible or reliable results (Hackshaw, 2008).

It has been well documented that small samples can produce false-positive results, or over-estimate the magnitude of an association (Hackshaw, 2008). Statistical power is the probability of rejecting the null hypothesis when it should be rejected. Many studies in psychology have low statistical power because of a low sample size, measuring small effects, or both. A study with low statistical power has a reduced chance of detecting a true effect. But what is less well appreciated, is that low power also reduces the likelihood that when a statistically significant result is found it reflects a true effect (Button et al., 2013). The consequences of this include overestimates of effect size and low reproducibility of results.

Big Data may Help the Replication Crisis

The replication (or reproducibility) crisis (term coined in the early 2010’s to add weight to the issue) has hit the fields of social psychology, educational, and developmental psychology the hardest (Open Science Collaboration, 2015). It highlights the finding that few psychology (and other natural sciences) studies can be replicated (Nosek et al., 2012; Open Science Collaboration, 2015; Zwaan et al., 2018). This can be in part attributed to (among other things such *p*-hacking, publication bias, and fraudulent results) small sample sizes. It may therefore be better to have a large dataset that can be interrogated thoroughly rather than a number of small studies that are hard to compare due to high sample variability.

Whilst reproducing the results requires using the same data, replicating involves collecting a new sample to test the effect (see Table 1.1 for an illustration of the differences between commonly used validity terms). Two large samples are less likely to have high sample variability as both are more likely to be representative of the true population than two small samples. For example, if sample A is large and sample B is small; if an effect is truly there, it is more likely to be detected in Sample A. If the effect is not truly there, sample B is more likely to erroneously detect it (assuming the same methods are being used, and there is no trade off of data quantity over quality). Therefore, a replication with a large sample (at least large enough to detect the reported effect size, calculated using a power analysis) is more likely to replicate an effect in A than B. Thus, even if the effect investigated is small, true effects are more likely to be found and replicated using large samples, as large samples increase statistical power. (This also has its drawbacks, which is discussed in more depth in Chapters 6 and 9). In many big datasets there is enough data for the authors to perform and report their own replication using a different randomly selected sample of the data (even if it is in the supplementary materials). This will help validate their results, and help reduce the problem of replications failing to be a) conducted, and b) published, due to the bias towards new and exciting research, and the difficulty of interpreting a failed replication (it doesn’t necessarily mean the effect isn’t there).

Table 1.1: A system for describing the concepts used to define the validity of a result

		Data	
		Same	Different
Methods	Same	Reproducibility	Replicability/ Model Generalisability
	Different	Robustness	Generalisability of Effect

Combining the Advantages of Field and Questionnaire Studies

In psychology, the trouble, cost, and time of getting participants into the lab for testing are responsible for driving sample sizes down. However, there are other research methods which inherently allow for the generation of data more easily. Field experiments, where an independent variable is manipulated in a naturalistic setting, are advantageous as they allow for the capture of behaviour in an ecologically valid setting (discussed further in Section 1.3.2. Although data can be collected on multiple people at once, the expense of having researchers ‘in the field’, entails that data is generally only collected for a small number of ‘windows’ in time and space. Additionally, due to privacy issues field experiments don’t allow for the ‘tracking’ (repeated measures) of an individual over time.

Questionnaire studies are cheaper and easier to administer, enabling the collection of more data. Whilst useful for measuring things of a subjective nature such as attitudes and beliefs, subjective well-being, or personality (all hard to be measured in a non-subjective manner as they require some degree of introspection), questionnaires measuring behaviour have been shown to be affected by a multitude of different things: the question order (although this can often be controlled for through randomisation) and interpretation, the structure of the rating scales, the anchoring heuristic (Tversky and Kahneman, 1986) (where question responses are effected by the response given to the previous question), leading questions (Loftus and Palmer, 1974), social desirability bias (Van de Mortel et al., 2008), or misremembering (Jakovcevic et al., 2014). In addition, sampling bias can occur due to the reliance on volunteer sampling. As a result of an accumulation of the above, the majority of the variance in the association between self-reported behaviour and objective behaviour remains unexplained (Kormos and Gifford, 2014). (This is discussed further in Section 1.3.2.) Big data measures real world behaviour objectively and fluidly through time and space. Data is already being collected for commercial (or other) purposes (which can have its drawbacks, as discussed in Chapter 9) and thus comparably minimal set up fee or effort is required. Thus, exploiting behavioural datasets is a logistically and financially accessible way for psychologists to gain objective records of behaviour ‘in the field’ *and* large sample sizes.

The Need for More Longitudinal Data in Psychology

Mass datasets can be harnessed to study much smaller effects which are not reliably detectable in small data (due to low statistical power). Novel, more subtle, associations can be

discovered, and tested for generalisability using out-of-sample testing or cross validation (see Section 2.1). Replicability of effects and generalisability of the model can also be tested for in different samples or new datasets of a variant type/quality. Not only will large samples enable greater statistical power to reliably detect small effects, the global nature and continuous stream of big data offers potential to drive down the expense of conducting longitudinal analyses, and studies comparing or sampling across multiple countries and cultures. There are a shortage of longitudinal studies conducted in psychology, limiting our ability to study humans and behaviour over time and as they develop in age (Kelloway and Francis, 2013). Those that do exist, often focus on ageing adults or childhood development and sample the same subjects at repetitious sparse intervals. Having large intervals is more practical (less measurements), but consequently some of the granularity is lost. Much can happen in a month of a person’s life, and this loss of information can affect the interpretation of the results in particularly, seasonal, or non-linear trends can be missed.

In his book *Outliers* (2008), Gladwell discusses an example whereby a closer analysis of the trends of school children’s learning patterns provides insights into the discrepancy between the learning rates of children from a lower social economic status (SES) and a higher SES. The following example uses this study to illustrate how more fine grained data over time can reveal new insights, and how incorrectly fitting a simple linear model could lead to different, even incorrect, conclusions. Measures of reading ability of school children in the US were acquired from when the children started school, to when they left the infants (after year 3). The child’s reading was measured once a year in September (after their summer holidays). Reading development was analysed by SES, and is illustrated in Figure 1.1 using fabricated data. From Figure 1.1a it appears that not only do children from a higher SES commence school with a better reading level, they also respond better to teaching and their reading develops at a faster rate. However, if we look at Figure 1.1b, which samples children’s reading ability over 3 points within the year (September, January, and July), the picture is different. The high SES group show a linear trajectory, whilst the low SES group develop at the same rate as the high SES group throughout the school year, but over the long summer holidays (between July and September) their development plateaus, whilst the higher SES group’s continues. During the summer holiday’s the teachers have no input into the children’s learning rate. The children from a lower SES are more likely to be spending their summer playing “out on the street” due to a lack of money for good child care. Whilst those from a higher SES group are “taken to museums or being read to” - spending their 7 week break in a more intellectually stimulating environment. This example illustrates how insufficient data and poor model fitting (i.e. without visually checking the shape/structure of the data) can lead to not only a mis-interpretation of the results, but also there is the potential for poor policy decisions to be made.

Policy and the allocation of resources are usually constructed with the help of data and statistical models. Data from Figure 1.1a would suggest the need to re-evaluate teacher time allocation or catchment areas; whilst the data from Figure 1.1b suggests time and money

would be better spent offering additional support to less well-off parents over the holiday period, or truncating the long summer holidays to minimise the growth of the ability gap. Some may argue that testing children more frequently can have a negative impact as it causes stress and forces the evaluation against peers from a young age, but data can be collected passively. Perhaps such data could be modelled on an individual level to identify different learning styles to personalise and optimise each child’s learning experience. This example highlights how fine grained temporal data might identify ways in which society can be fairer and reduce discrimination against its less fortunate individuals. In this case, the insights gained from such ‘big data’ might enable societal structures and state institutions to help children’s development be more independent of SES.

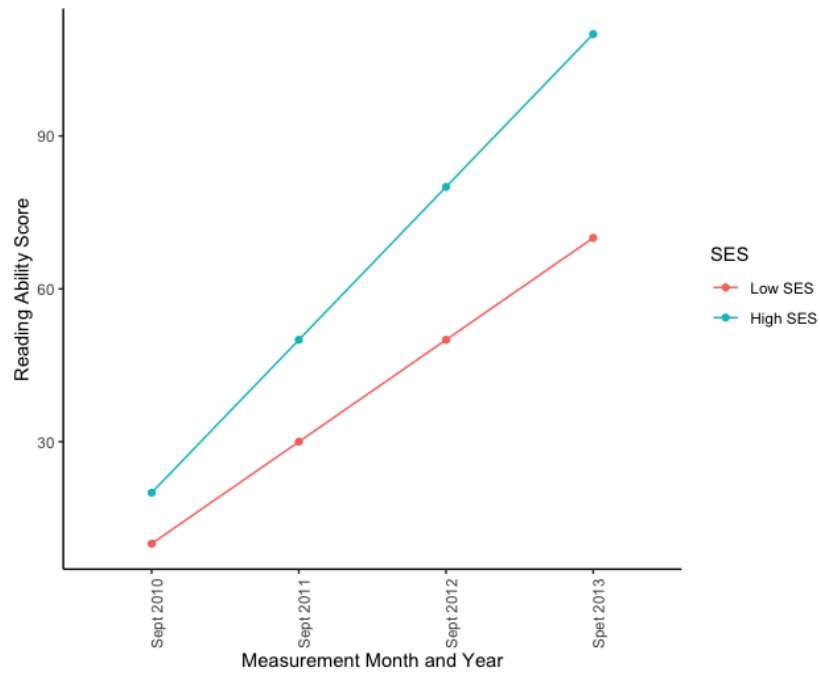
Unconstrained Analytic Approaches

A final advantage inherent to the scale of big data, is the ability to apply more complex analytical techniques (such as machine learning methods) to the data. Machine learning algorithms can detect complex patterns within the data. However, it is the size of the data that allows algorithms to capitalise on using multiple variables, and allows for the modelling of complex interactions and non-linear effects. With greater complexity of models comes an increased likelihood of the algorithm modelling noise in the dataset (over fitting). This is mitigated against by using techniques such as out of sampling testing and cross validation. These validation methods and machine learning models are discussed in more detail in Section 2.1.

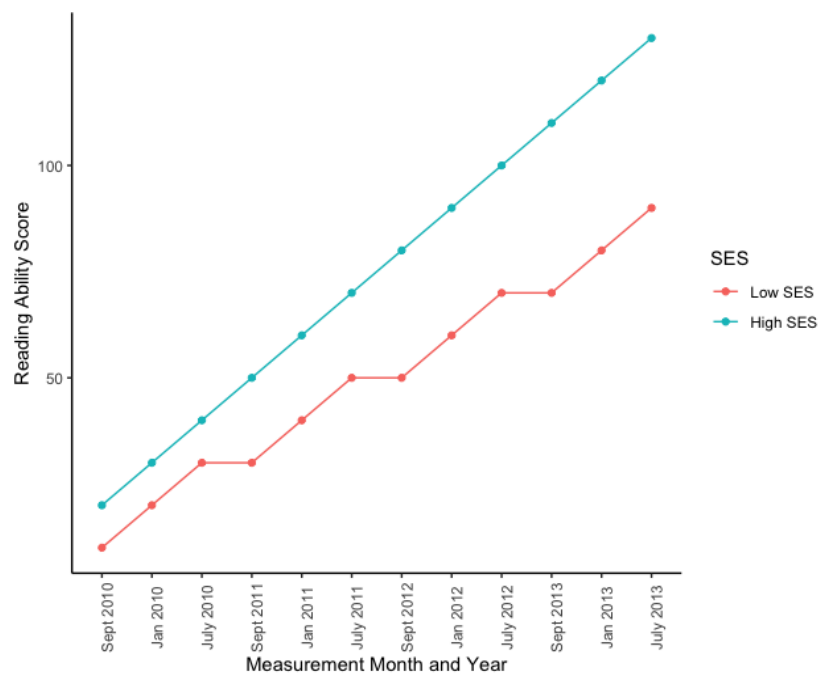
As well as the size of the data, the variety of the data now available to use also facilitates novel analyses to be undergone. For example, geotagging (data being labelled with a standardised location code or coordinates) big data allows the analysis of spatial or geographic effects - something that is hard to study with small amounts of data because of high sparsity. In addition, data being collected as a continuous stream with time-date stamps allows for a more intricate look at how behaviour develops over time. The temporal and spacial component can be combined to study patterns of behaviour, such as migration, over both space and time. All in all, the large quantities of big data allows for the use of more complex and powerful algorithms that can learn subtle and more convoluted patterns than the methods that can be applied to small data.

1.3.2 Naturalistic Behaviour Capture

A second area of challenges facing psychology lies in the goal of capturing ‘real’ behaviours and views. In psychology, real world behaviour is hard to measure, yet it is vital if research is being used to influence policy or decisions made in the real world. Lab based experiments, although useful at studying an isolated behaviour in controlled conditions, lack ecological validity, i.e. they don’t necessarily tell us about how such effects will generalise outside of the lab. Field studies, however, measure real world behaviour by observation and interviewing



(a) Yearly sampling



(b) Sampling 3 times a year

Figure 1.1: Two examples of data sampling using fabricated data. Plot a) shows sampling of children's reading ability once a year in September. Plot b) shows sampling 3 times in a year in September, January, and July

in a natural setting. Field studies still involve a manipulation of an independent variable, but ecological validity is retained as the setting of the study is ‘everyday life’. Researchers conducting field experiments usually observe individuals whilst taking notes, and will often ask questions at the end.

Demand characteristics refers to an experimental artefact which occurs when subjects subconsciously alter their behaviour because of actually knowing or thinking they know the purpose of the study and how the researcher expects or wants them to behave. Demand characteristic are less of an issue with field experiments than lab experiments as participants don’t always know they are being observed. If subjects are unaware they are being analysed, behavioural measures produce a more natural and accurate measure of every day behaviour. Yet, if individuals are being observed unawares, issues of informed consent arise. If participants are not made aware of their participation in an experiment, privacy must be respected during observations and therefore personal or identifying information cannot be collected.

Another way real world behaviour can be measured is using surveys. Surveys can ask participants to report on their behaviour over the last weeks/months/years. A problem with this method, however, is that self-reported behavioural measures can be affected by social desirability bias (Van de Mortel et al., 2008) (reporting behaviours more in line with the social norms either consciously or unconsciously to appear more favourably) or misremembering (Jakovcevic et al., 2014).

For example, in health psychology, getting an accurate measure of calorie and food group intake is a challenge. Both physical and psychological characteristics of study participants have been shown to affect the reporting bias (Westerterp and Goris, 2002). A review of the literature by Westerterp and Goris (2002) found that the degree of misreporting calorie intake can increase with repeated assessments for the same subject, confounding the results of intervention studies. In addition, demand characteristics drive subjects to report according to expected instead of real nutritional intake; and in clinical settings, an increased awareness of nursing staff lead participants to over-report consumption. Human memory and even perception is fallible, and been shown to be affected by mood (Bower, 1981), stress and anxiety (Roozendaal et al., 2009), context (Smith, 1994), and leading questions (Loftus and Palmer, 1974). In the field of environmental psychology a meta-analysis found that 79% of the variance in the association between self-reported behaviour and objective behavioural measurements was unexplained (Kormos and Gifford, 2014). Such a discrepancy is troubling when real world interventions are being devised on the results of studies using self-reported behaviour. If actual behaviour can be measured accurately, on a large scale, more accurate insights into different behaviours can be made, and used to inform future policy decisions.

One of the defining features of big data is that it is already being collected for non-research purposes, unlike small data which is generally collected with a particular research hypothesis in mind. This means that the observed behaviour is recorded in the data, and the data being collected and stored, whether the research is being carried out or not. The advantage of this (there are also disadvantages, which are discussed in Chapter 9) is an individual can not be

aware if their data is being collected for a specific hypothesis, or what that hypothesis may be. Accordingly, the individual cannot change their behaviour in response to a hypothesis, even if they wanted to influence the results in a particular direction. Supposing a few individuals were aware of such a study being conducted, and wanted to sabotage the results, they would need to persuade large numbers of accomplices to alter their behaviour, to make even the smallest ripple of disturbance in the ocean of big data.

Generally, people are aware they have given company X permission to store and analyse their data, most likely for marketing reasons, when they accept the terms and conditions during the sign-up/registration. Although, too often people don't realise their data is being collected and analysed at all (Vizard, 2018). The data being collected is just a normal part of people's every day life. They don't see anyone watching or analysing them. This de-personalisation, lack of physical proximity to the people analysing the data, and the integration of data into their daily lives all contribute to the majority of the population not thinking or caring too much about the data they are producing (Vizard, 2018). This all makes big data an unobtrusive way to measure human behaviour in the real world, ensuring it is not confounded by demand characteristics.

1.3.3 Studying the Underrepresented

In small data, effects are usually generalised across the whole sample (or a couple of control variable dimensions) attributing one effect size, nature, and direction to the population. In the instances when data is analysed by a control variable, for example gender, this halves the sample size and reduces the statistical power for finding within subsample effects. Therefore, with small data, broad sub-groups can often only be analysed by being split on a few of key variables, or only a few interactions analysed, before the sub-samples are too small or the data too wide (too many variables compared with data points) to produce reliable results.

For example, as most psychological research uses psychology undergraduates, let's imagine the majority of people in an American psychology department are White, 10% are Asian, 7% Black or African American, 2% American Indian, and 2% 'Other'. If the sample size of a study is 50 people, the chances are you will only have 5 Asian subjects, a few Black or African American, and you may not study a single American Indian. Under such conditions it is impossible to study whether your observed affect is confounded or dependent upon race. This creates a massive bias as studies generalise the effects present in the sample across all subgroups, and making sweeping conclusions. Furthermore, the well documented Simpson's Paradox (Simpson, 1951) describes cases where an effect can not just be different within subgroups but even reversed, compared with the main effects across the whole sample. Figure 1.2 illustrates the Simpson's Paradox on fabricated data of an imagined example. Both graphs have the same positive correlation between income and well-being. The graph on the right however takes into account the effect of country, and shows that within a country the relationship is actually reversed, where the more money someone has the less happy they are

- but some countries are happier and wealthier than others. The Simpson’s Paradox is an example of why it is dangerous to assume the main effects are the same as the effects within subgroups. The smaller the subgroup, the lower their chance of being explicitly studied, and the greater the chance of being discriminated against as actions and decisions are based on what is best for the majority.

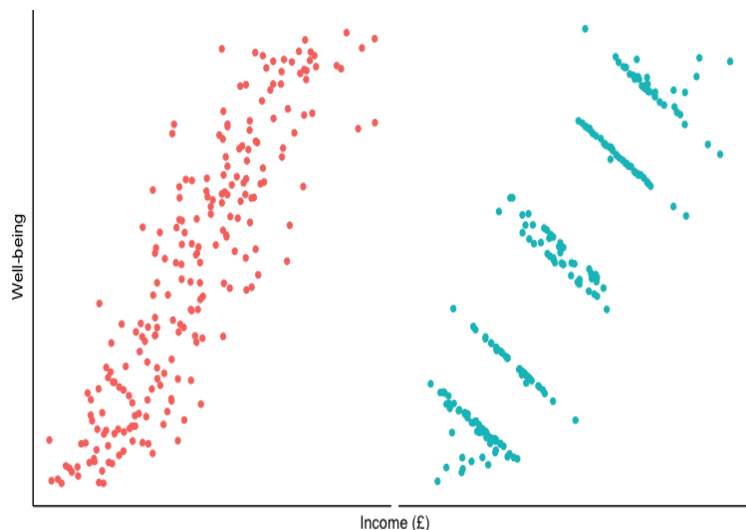


Figure 1.2: Example of the Simpson’s Paradox on fabricated data. The graphs show Income on the x axis, and Well-being on the y axis.

In smaller studies subgroups can be ignored, or even have their data thrown away as ‘outliers’. Big data allows the identification and comparison of effects in different societal sub-populations, that would otherwise be too small to identify and statistically study in small data. This can improve our knowledge and understanding of how culture, religion, and social class affects behaviour. Not only can subgroup analysis be conducted, but those groups considered to be outliers, minorities, or marginalised can explicitly become the focus and studied in their own right. Of course, this requires that explicit labels exist in the data of the subgroup of target, in order for them to be identified, and it remains an ongoing challenge to conduct such analyses using inferred or unreliable labels. But the process of knowledge discovery could be reversed, and irregular behaviours identified from the bottom-up. These can later be interpreted/explained using the meta-data (data about the data) or existing data labels. People can be grouped not just by demographic labels, but by their behavioural patterns. For example, people who work at night can be identified through wearable tech, GPS, or heart rate data; veganism can be inferred from transaction data; and people who are in debt can be recognised using mobile money apps or credit ratings. In these cases the labels are inferred directly from the data, creating the opportunity to analyse subgroups of people without having to questionnaire or survey them. Overall, big data is an exciting new way to reveal new insights about ‘the few’ as well as ‘the many’.

A Bias to WEIRD Samples

The majority of psychology research is conducted in European or American universities. This creates an over reliance on WEIRD populations (Western, Educated, Industrialised,

Rich and Democratic) (Henrich et al., 2010). A survey of the top psychology journals found that 96% of subjects were from Western industrialised countries — which house just 12% of the world’s population (Arnett, 2008). Despite this, research routinely assumes that their results are broadly generalisable. A growing amount of evidence shows that basic cognitive biases and processes significantly vary across different populations and cultures. For example, Americans, Canadians and western Europeans rely on analytical reasoning strategies — which abstract objects from their contexts and use rules to explain and predict behaviour — substantially more than non-Westerners (Henrich et al., 2010). Conversely, Asians tend to reason more holistically, taking into account situational influences when analysing people’s behaviour. Despite the vast and growing Asian populations, long-standing theories of how humans perceive, analyse, and remember still focus on analytical thought.

When researchers are looking for study participants, undergraduates, particularly psychology undergraduates, are an easy target. Physical proximity of classes and lectures makes them more likely to respond to poster advertisements; they are easily contactable through mailing lists; being students, they are more likely to be responsive to the small amounts of inconvenience allowance on offer; and finally, many undergraduate programs make it compulsory for students to partake in a certain number of studies to pass the academic year (using a system of credit accumulation). All this amounts to the psychology undergraduate population in western universities being the most rigorously tested and studied population of individuals. Again this situation is precipitated by the bias towards actively collected data, rather than passive secondary sources.

Undergraduates are typically in their late teenage years to early 20’s, therefore their brain is not fully developed (Johnson et al., 2009). Studies have uncovered many differences between undergraduates and a non undergraduate population, including differences in learning styles (Stallman, 2010), a higher prevalence of mental health issues (Stallman, 2010), and differences in the need for cognition and preferences for competitive or cooperative strategies (Ward, 1993) to name a few. Therefore, there is a major limitation with studies using undergraduates, in terms of how well the results generalise to a non-undergraduate population.

Big data has potential to help remove sampling bias by being able to shed light on the behaviour of a greater number of individuals from a greater variety of countries and cultures. Even within the poorest of countries, much of the population who would usually be out of the reach of researchers, have smart phones. In 2017, 97.8% of adults in the developing world carried mobile phones (Sharwood, 2017) which can be used for calls, messages, mobile banking, maps and apps - all of which produce streams of potentially insightful data. These streams of data can allow researchers to analyse and understand a greater diversity of people, and creates the opportunity for more cross-cultural comparisons to be made.

1.3.4 Preventing Overfitting

Another phenomenon thought to contribute to why the rate of successful replications in psychology is low, is that of overfitting (Yarkoni and Westfall, 2017). Overfitting is where a model fits well to a given sample, but is too finely tuned to that sample data to be able to generalise to the wider population. If models are built to fit the researcher’s data too closely, they cannot be replicated by others on new data (Open Science Collaboration, 2015). This stems from an explanation-only approach used predominantly in psychology and discussed in more detail in Section 1.4. In psychology, models are predominantly evaluated by the amount of variance they explain in the sample. The model is fitted to the sample, and the residual variance is measured using the R^2 metric. Yarkoni and Westfall (2017) demonstrate that when using the R^2 metric for variance explained on the sample data, or even adjusted R^2 (where the metric is adjusted to penalise large numbers of variables in the model), this is an overestimate of what the R^2 would be if the model was used on a novel, unseen dataset. Thus by not taking into account how well a model generalises to new data, effect sizes are being exaggerated.

The overriding reason why psychologists don’t test out-of-sample or use data recycling methods like cross validation (discussed in more detail in Section 2.1) is because researchers don’t have spare or extra data - they need all the data they can acquire to increase the statistical power in their studies, to have a chance of finding the effect they are investigating. Acquiring enough data to be able to hold some back allows prediction models to be tested on how well they can generalise out-of-sample, and thus reducing type 1 errors (false positives) caused by over-fitting. By using big data samples, models can be checked for generalisability, and fewer false positive results will be published - in turn increasing reproducibility and replicability.

1.4 Prediction versus Explanation

The previous section outlined the advantages of big data, yet it is important to note that big data requires a different approach to science which will be outlined in the following two sections discussing a prediction versus explanation approach, and an inductive versus deductive approach to science. Big data and machine learning come hand in hand because such powerful prediction algorithms are only useful when large datasets are available to train on. The field of machine learning is typically preoccupied with prediction, i.e. building algorithms and models to predict an outcome with as high accuracy as possible. As touched upon in Section 2.1, higher accuracy in big data usually (but not always) requires greater complexity in a model, which comes at the cost of interpretability. The field of psychology comes from the opposite angle, focusing on explanation first. An explanation approach, requires utilising ‘simple enough’ models which can then be primarily *understood*, sometimes at the cost of accuracy. Therefore, conceptually how a computer scientist and a psychologist typically view a research question using big data differs dramatically.

A psychologist’s goal is to understand human behaviour, whilst a computer scientist tends to want to predict it. The two goals are related but not synonymous. Understanding human behaviour can encompass both explaining it and predicting it, although these two goals are not equivalent. What good is an explanation that can not predict anything, and vice versa? Prediction involves finding the best model to forecast future observations. It is possible to make a good prediction model that is causally implausible (Mahmoodi et al., 2017), whilst explanation requires identifying the causal processes of what gives rise to observed phenomena (Mahmoodi et al., 2017). For example, a psychologist wanting to understand which personality traits affect well-being, being able to predict well-being accurately using a black box algorithm, tells them that personality is important, but nothing helpful for understanding which traits are important and why.

Typically, psychologists and computer scientists therefore utilise different tools and methods. Psychologists, and most others in the social sciences, fit a model (i.e. a regression) to the sample, evaluating the ‘goodness of fit’, or how well the model can explain the variance within that sample. Then, the parameter values (i.e. regression coefficients) can be analysed for statistical significance. In an explanation based approach there is only a measure of how well the behaviour in the sample can be explained. If the sample is small, or not representative of the general population, the results will not be either. Whilst a prediction approach involves testing a model on data out of the sample it was trained on (see Section 2.1). The model’s performance is evaluated in terms accuracy, compared to other models or a baseline model. Typically statistical uncertainty is not quantified in machine learning (Faraway and Augustin, 2018; Strobl and Zeileis, 2008).

The goal for psychology is to understand human behaviour, and despite psychology predominantly using an explanation only approach, understanding involves both being able to explain and predict. It is not useful to explain behaviour if we cannot do anything with that knowledge. If we truly understand a behaviour we should be able to test that understanding by predicting it. In addition, we need to be able to predict behaviour in order to offer the right interventions, and try to change it. On the flip side, we want to explain behaviour, through the development of theories, in order to be able to build better prediction models incorporating theory. Therefore, psychologists could benefit from using a combination of both explanation and prediction approaches.

Explanation and prediction require different methodologies and models - and hence tend to be used to answer different questions (Shmeli, 2010). Prediction models are more likely to answer questions which are applied and need a practical solution; whilst if the goal is knowledge/theory, an explanation approach may be more suited and allow cross-domain application. With the emergence of big data, and the commercial advantages that can be gained from accurate predictions of behaviour highly complex and impermeable ‘black box’ models have been the dominate approach to big data.

1.5 Induction versus Deduction

As well as the differences in prediction/explanation, another cultural difference between psychology and the field of machine learning is the approach to how data and hypotheses should be used in the knowledge discovery process. The debate between induction and deduction is much wider than in the context of big data psychology, yet is something which inherently emerges when the two fields meet. The two approaches, along with their advantages and disadvantages, will therefore be explored throughout this thesis.

One of the advantages of working with big data is that by applying complex methods of analysis, patterns can be unearthed which are not recognisable in smaller datasets. The nature of this pattern recognition is inductive, guided by the data. The scientific practice recognised in psychology, however, is deductive (Popper, 1949). Deduction starts with a theory-driven hypothesis, and data is collected to test it (top-down approach). Traditionally an exploratory approach of ‘lets see what is in the data’ doesn’t sit well with the discipline of psychology. This may be in part due to the issue of *p*-hacking. *P*-hacking (also known as data dredging, data fishing, data snooping, and data butchery) is the manipulation of the analytic or data collection process that increases the chances of finding a statistically significant finding. Examples of *p*-hacking include: performing multiple statistical tests on data, and only reporting the significant ones; stopping data collection early when a significant effect has been found; identifying a trend and then continuing to collect more data until it reaches significance; or removing ‘outliers’ without adequate justification to get a significant effect. Greater awareness of the prevalence of *p*-hacking has caused a movement towards pre-registering hypotheses, analytic methods/models, sample sizes, and expected effect sizes before commencing data collection - all with the sound goal of promoting generalisability. The inductive approach, however, doesn’t allow for the pre-selection of models. Instead, after trying out different models, the model that offers the best prediction is selected. While it is clear why many psychologists instinctively react towards pre-registration, machine learning also focuses on generalisability - just via different methods (this is discussed in more detail in Chapter 6, Section 6.1.2). Furthermore, effects sizes in small data are not comparable to big data and not typically produced by machine learning models at all. This makes pre-registering almost impossible using an inductive approach. Thus, with all the warnings about *p*-hacking in the psychology community, it isn’t any wonder that researchers are reluctant to adopt a “look at the data” approach to research.

A limitation with a deductive approach is that it requires researchers to themselves come up with a theory to test. It is impractical and perhaps unrealistic to expect researchers to always have eureka moments unprovoked by data, and to not expect biases in the theories they formulate. Particularly in fields or domains where there isn’t much past literature, a deduction only approach might even hinder the scientific discovery process (especially if we are waiting for scientists to be enlightened by an unprompted theory). Further, social pressures mean deduction can also threaten to undermine the creation of robust theories,

as researchers can feel the need to pretend they already knew in advance what was later revealed by the data (Kerr, 1998), otherwise known as HARKing: hypothesizing after results are known.

Further, using a deductive only method, a researcher is more invested in rejecting a specific null hypothesis. This can increase the risk, either intentionally or subconsciously, that the researcher interprets ambiguity in the alternative hypothesis' favour, falling victim to confirmation bias (Nickerson, 1998), or resorting to *p*-hacking. Focusing on a few specific hypotheses in a deductive approach can also prevent further exploration or analyses that could lead to future knowledge discovery. When using an inductive approach, any findings can be reported, whether it was resulting from a planned investigation or not. This allows researchers to be less constrained and therefore perhaps more likely to make a useful contribution of knowledge.

Nevertheless, if there isn't data available, and data collection requires the use of resources (i.e. time and money), then it is most efficient that researchers collect data that will try to answer their specific question. In this instance, they must need to know what hypothesis to test, and use theory to work out how to test it, all prior to data collection. However, in an age where data is being continuously produced, collected and stored, this logic becomes obsolete. Previously, when only small to moderate datasets have been available, induction has been less reliable. Now, effects can be detected in mass datasets reliably and even if the *a priori* probability (calculated by logically examining the existing information) is low (Mahmoodi et al., 2017). Therefore, if psychology does not reassess its approach to knowledge discovery, it is limiting the opportunity for novel insights to be discovered by adhering to a deduction only approach.

1.6 Potential Barriers to Big Data Psychology

Above, the introduction has made the case that big data and machine learning have the potential to augment psychology. So why isn't the use of such tools in psychology more prevalent? A growing number of psychologists are indeed using big data in their research over a widening number of topics (for more see Chapter 2). However, psychologists capitalising on big data for research are still few and far between (and certainly compared to researchers from other disciplines such as the hard sciences). Because the majority of researchers studying behaviour in big data are predominantly not from psychology, this often means they will lack the theoretical knowledge underpinning the topic. And therein lies the danger that researchers will miss or overlook the theoretical implications of findings. Researchers from a non-social science background may also lack the ethical training or standards accompanying research with human subjects (Buchanan et al., 2011) leading to oversight on ethical considerations. Psychologists are needed here, but research via big data, requires many to overcome barriers. As well as the conceptual shifts involving shifting between explanation and prediction, and from top-down deduction to bottom-up induction,

other, more practical issues such as the language, statistics, and technical skills used, can prevent psychologists from fully exploiting mass datasets.

For many psychologists, a skills gap perhaps also exists. This often starts with less coding or programming skills being prevalent in the field. Professors are often under a lot of pressure to get publications, and already struggle to find the time for their own research among teaching commitments. Finding the time to learn a new skill, such as programming, is highly challenging (and may also be viewed as an inefficient use of time, if the end goal is publications). In many institutions, psychology undergraduates are also not taught data analysis in a coding language such as python or R - instead using interface programs like Excel or SPSS.

Although a vital part of it, the skills gap does not of course entirely consist of programming. Anyone working with big data also requires an awareness of the type of models or methods which can and should be applied to big data. Fundamentally, big data should not be treated the same as data obtained from a lab experiment. In a lab, the idea is that everything (or as much as possible) is controlled - with the exception of the variable being tested. Thus, the effects found can be attributed to the independent variable in question. In big data, however, an infinite number of different variables, whether measured, latent or unknown, could be influencing behaviour. The more features that can be measured and input into the model, the better chance the model has of detecting patterns in the data. However, if all the features are input into a linear regression model, without accounting for multicollinearity, incorrect interpretations can be made from the fit coefficients (Minitab, 2013). This is discussed in more detail in Chapter 7. In addition to multicollinearity affecting results, if the model is not tested on unseen data, incorrect conclusions about the model's generalisability can be made (Yarkoni and Westfall, 2017). Thus, if big data is procedurally and statistically treated the same as small data, unreliable results will likely emerge. A small data researcher cannot simply transfer their understanding and methods of statistics over to big data, thus future practitioners would likely benefit from a formal education on the research methods and statistical properties of big data.

A further obstacle, arising from psychological research traditionally being occupied with explanation, is the foreign language and differing use of statistics being utilised by the prediction camp. Machine learning models are evaluated using prediction accuracy or classification error. This includes using metrics like precision, recall, F1 score, Area Under Curve (AUC), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). For a psychologist who primarily deals with effect sizes, β/b coefficients, correlation coefficient r , F -stats, t -values, p -values and R^2 values, this may be an adjustment. The differing metrics used also make it harder to compare effects and models across disciplines, making formal comparisons such as meta-analyses difficult.

In psychology, an effect is not generally considered, unless it is significant. The significance of effect sizes are evaluated using p -values. 'Results' which don't meet the significance criteria (most commonly set at $p < 0.05$) are barely reported, let alone published as a finding.

Whilst, for machine learning methods, significance is not the gold standard. In prediction analyses, there are no cut-off values for when a model is useful (other than relative to a baseline) or whether a variable is important or not. When millions of data points exist, random variables could be significant either by chance, sampling, or the result of some unmeasured latent factor. Further, when the best possible prediction is the goal, a small increase in prediction power (even if it is not significant) is still useful. When a small increase is applied across millions of data points, the net effect can still be large.

In summary, for psychologists to exploit big data for psychological research, they may need to break away from previously learnt traditional practices and will potentially have to learn new methods, models, statistics, programming languages, and a new approach to knowledge discovery. This thesis hopes to contribute somewhat to this journey, by critically engaging the reader with such big data methodologies.

1.7 Overview of the Thesis

The introduction has argued that there might be a role big data/machine learning techniques have in supplementing and potentially bolstering psychological research. This thesis exploits large transaction data connected to a mass survey, to predict and explain consumer behaviour in big data. This thesis will, through analyses in three different domains (well-being, the environment, and mental health), demonstrate the advantages of using big data for psychological research. This work will also navigate through the issues which can arise when attempting to combine machine learning with psychological research practice, particularly when trying to appease both the need for accurate prediction and theoretical explanation.

The next chapter, Chapter 2, is in two parts. The first introduces machine learning in more detail, and the methods used in this thesis. The second half defines the sphere *Big Data Psychology* in more detail, the aims of such an approach, and reviews the previous relevant literature. The literature highlights that most of the work is being done in the domain of consumer/marketing psychology, driven by commercial goals, and primarily by researchers not from within the field of psychology. This thesis aims to use big data to make theoretical contributions to the fields of (mental) health and environmental psychology. Hence, this thesis consist of three core research areas: well-being, single-use plastic bag purchasing, and anxiety.

Chapter 3, the methodology chapter, first walks the reader through the literature used to construct the mass-distributed personality questionnaire that was sent to loyalty card users. The survey is a crucial element of the research presented in this thesis, and what makes it possible to conduct psychological research of this nature using big data. The review outlines how different psychological traits can affect how people shop. The important psychological traits for predicting consumer behaviour in the literature were measured in the questionnaire. The second part of this chapter describes the main datasets which are used in this research. It describes the transaction data attained from all loyalty card users across the UK over

10 years, outlines the survey questionnaire data and the distribution of the psychological variables within the data. Lastly, it describes the transaction data for the respondents of the questionnaire.

The first of five analytical chapters, Chapter 4, focuses on predicting subjective well-being (SWB) through the products and brands people purchase. Through predicting well-being, a better understanding can be gained of what people should spend their money on to make them happier. The first study is a replication (same methods, different data) and test of the generalisability (different data, different methods) of a study by Matz et al. (2016) which uses the similarity of one’s personality to their shopping’s personality to predict SWB. The second study refines a new method of using the data to infer products and brands’ personalities. It conceptualises a novel data-derived metric as being akin to the principles of recommender systems - “buying what people like you buy, makes you happier”.

Chapter 5 extends the research in the previous chapter. The regression model that was used in Chapter 4 is tested against ‘out of the box’ machine learning algorithms, introducing the method of cross validation to reduce overfitting. The data is also analysed using decision trees and random forests, and it is argued that these can be used to gain further insights into the structure of the data, the non-linearity’s, and the interactions between independent variables.

The third analytical chapter, Chapter 6, investigates which psychological and socio-demographic variables predict frequent plastic bag buying after the 5p levy was introduced to the UK. The methodology utilises an inductive algorithmic modelling approach to research, where different models are optimised on training data, and the ‘best’ model is selected. The work then exploits the detailed transaction data further to refine the research question, moving from predicting how many plastic bags someone buys over a time period, to how many bags given the number of shopping trips it was estimated a bag was *needed*. Using this question, a methodology is honed for statistically extracting the significant predictors from a logistic regression model fit to big data, whilst protecting against the notion that both the model and the significance of the predictors could be effected by noise.

Chapter 7 discusses the problems of multicollinearity in big data, particularly for interpretation, as found in Chapter 6. The chapter then explores other models/methods, namely decision trees, random forests, and Non-negative Matrix Factorisation (NMF) to mitigate this problem. The analyses supplement the variable importance insights gained from regression in the previous chapter (when multicollinearity is suspected to be interfering with the results) predicting plastic bag consumption.

The final experimental chapter, Chapter 8, looks at whether the purchasing of over the counter anxiety medication, could be an unobtrusive objective measure of anxiety levels in the population. This is done by understanding the relationship between the behaviour and the self-reported personality traits which have been found to predict anxiety in the literature. The strength of the signal in the data is analysed, as well as the predictors, to try to understand the behaviour better, and how it may be qualitatively different from anxiety

symptoms measured using self-report questionnaires. Finally, exploratory investigations into the geographic trends of personality and anti-anxiety purchases are conducted to illustrate how geo-tagging, often present in big data, can allow the linking of datasets to other open source data.

Finally, Chapter 9, the Discussion, reviews how the research in this thesis has contributed to the discussions outlined in this introduction. The advantages and limitations of big data are discussed, as well as the ethics of big data and big data psychology research. Finally, frameworks are proposed for how research in big data psychology can be conducted using both prediction (to prevent overfitting) and explanation (for theoretical insights); as well as using both inductive and deductive methods for hypothesis generation and testing. This chapter finishes with concluding remarks on the utility of big data in psychology for both psychology, and the field of machine learning.

Chapter 2

Background and Literature Review

As outlined in the Introduction, one of the proposed advantages of big data is it allows for the use of more complex machine learning algorithms, which have the power to detect more subtle effects, interactions, and non-linearities. This chapter introduces machine learning as a field, as well as the methods used in this thesis, namely decision trees, random forests, and cross validation. It outlines the potential advantages and disadvantages of using decision trees and random forests, and why cross validation is needed in big data. Following this, in the second half (Section 2.2), the sphere referred to as big data psychology is defined, and the goals for what could be achieved by combining machine learning with psychology are outlined. Lastly, a review of the literature relating to big data psychology is carried out and the gaps which this thesis hopes to address, are identified.

2.1 What is Machine learning?

Statistical machine learning, originating in computer science, is now utilised across a broad span of disciplines. Machine learning methods consist of a group of powerful prediction techniques where the algorithm can learn from a training dataset, and then make a prediction on a different, novel dataset. Machine learning is a bottom-up discovery process where algorithms detect patterns which are tested to see if they generalise. There are two classes of machine learning algorithms: supervised and unsupervised. Most methods used are supervised, where the algorithm is told what variable to predict (the Dependent Variable (DV)) and the possible Independent Variables (IVs) it can use to do so. The model then has to learn the how to use the IVs to best predict the DV. Unsupervised learning is when there is no outcome variable to predict - only independent variables. This method can be used to model the underlying structure or distribution of the data, to reduce the number of features, or to select clusters. Emergent structures can also be tested on unseen data to check that the algorithm has not overfitted to the sample. The most common split of the data is 80% used to train an algorithm, and 20% to test it. However, this can vary depending on the amount of data and the research question.

Some machine learning methods lend themselves better to interpretation than others. Generally, the more complex a model is, the harder it is to interpret. For example, stan-

standardised weights on a simple linear regression make it easy to see how much each variable contributes and in what direction to the final prediction. However, the world is not full of nice simple linear relationships. Variables interact, and mediate and moderate each other, creating non-linearities and subgroups. Thus, the more complex models that can learn these complexities often perform better. This usually comes at the cost of being harder to interpret. Deep Learning methods such as artificial Neural Networks (NN), can be so complex that humans often don't understand how they are working, they are "black box" algorithms. The NN constructs predictive features that can be so highly dimensional and abstracted that humans cannot see or comprehend them. There are however some methods which produce a compromise between the simple linear regression coefficients and abstract high dimensional features. Decision trees and random forests are powerful non-linear algorithms, yet still produce interpretable variable importance metrics. The remainder of this section introduces the methods decision trees, random forests, and cross validation.

2.1.1 Why use a Machine Learning Approach?

To most in scientific psychology, the goal of understanding human behaviour involves being able to explain behaviour, and to predict it. Psychology research, however, has traditionally focused on the former, prioritising explanatory investigation. This is perhaps due to the primary positioning of causal mechanisms in social sciences, with methods of prediction taking a secondary role (Yarkoni and Westfall, 2017). While related, prediction and explanation are not synonymous (Shmueli et al., 2010) and recent work has highlighted that an explanation-based approach only may generate increased false positive results or inflated effect sizes (Yarkoni and Westfall, 2017). These issues can be explained by the phenomenon of 'overfitting', where a model fits well to a given sample, but is too finely tuned to the noise within that sample's data for it to successfully generalise to the wider population. Indeed, Yarkoni and Westfall (2017) suggested that this may be a contributing factor as to why the rate of successful replications in psychology is relatively low - models are being built to fit researchers' data too tightly, generating greater explanation of a specific sample but hindering generalizability, and thus reproducibility by others on new data (Open Science Collaboration, 2015). Yarkoni and Westfall (2017) demonstrate that when using the R^2 metric for variance explained on the sample data, or even adjusted R^2 , this is an overestimate of what the R^2 would be if the model was used on a novel unseen dataset.

In recent years, advances in statistical machine learning approaches, which focus on predictive rather than explanatory goals, have opened the doors to new tools for expressing relationships between covariates. A group of powerful prediction techniques are now available where algorithms focus on fitting training datasets in a manner that makes optimal predictions on external data. Because models are validated on novel data, machine learning techniques can help researchers to avoid overfitting, and thus potentially reduce false positive findings. Yet, because psychologists also want to explain relationships making model

interpretation (i.e. understanding what variables the model is using to make the prediction) crucial. Given this, decision trees and random forests can provide a compromise between powerful prediction algorithms, and high interpretability.

2.1.2 Decision Trees

Decision trees refer to both regression and classification trees. Regression trees make continuous predictions, whilst classification trees make classification predictions (either binary or multi-class predictions). They recursively split the feature space (in two dimensions this creates rectangle shapes, see Figure 2.2) to create a tree-like structure with decision points and branches (see Figure 2.1. Each node (represented as a bubble in the diagram in Figure 2.1) represents a sub sample of the data. Decision trees split the data using binary decisions in a way which reduces the most variance within the data or subsample (Breiman et al., 1984). The splitting process is repeated on each new subsample, creating the tree structure.

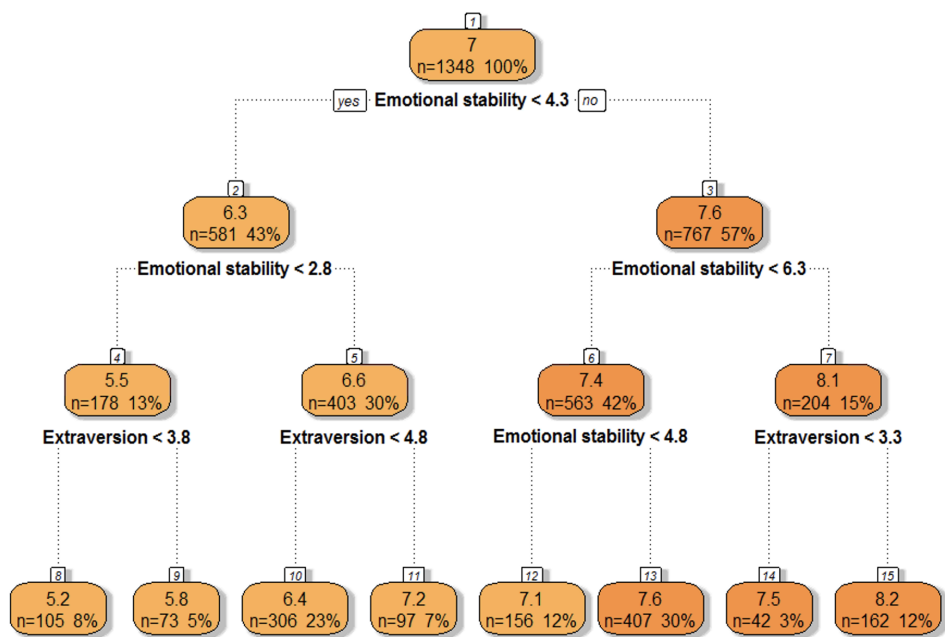


Figure 2.1: A simple decision tree using Extraversion and Neuroticism to predict Well-being. This tree uses 7 splits, to create 8 terminal nodes (along the bottom). In each node there is a prediction of that subgroups well-being score (top, middle number), the number of people in each sub-sample (bottom left) and the number as a % of the total sample (bottom right). The lowest well-being prediction this tree makes is 5.2/10 (node 8), and the highest is 8.2 (node 15).

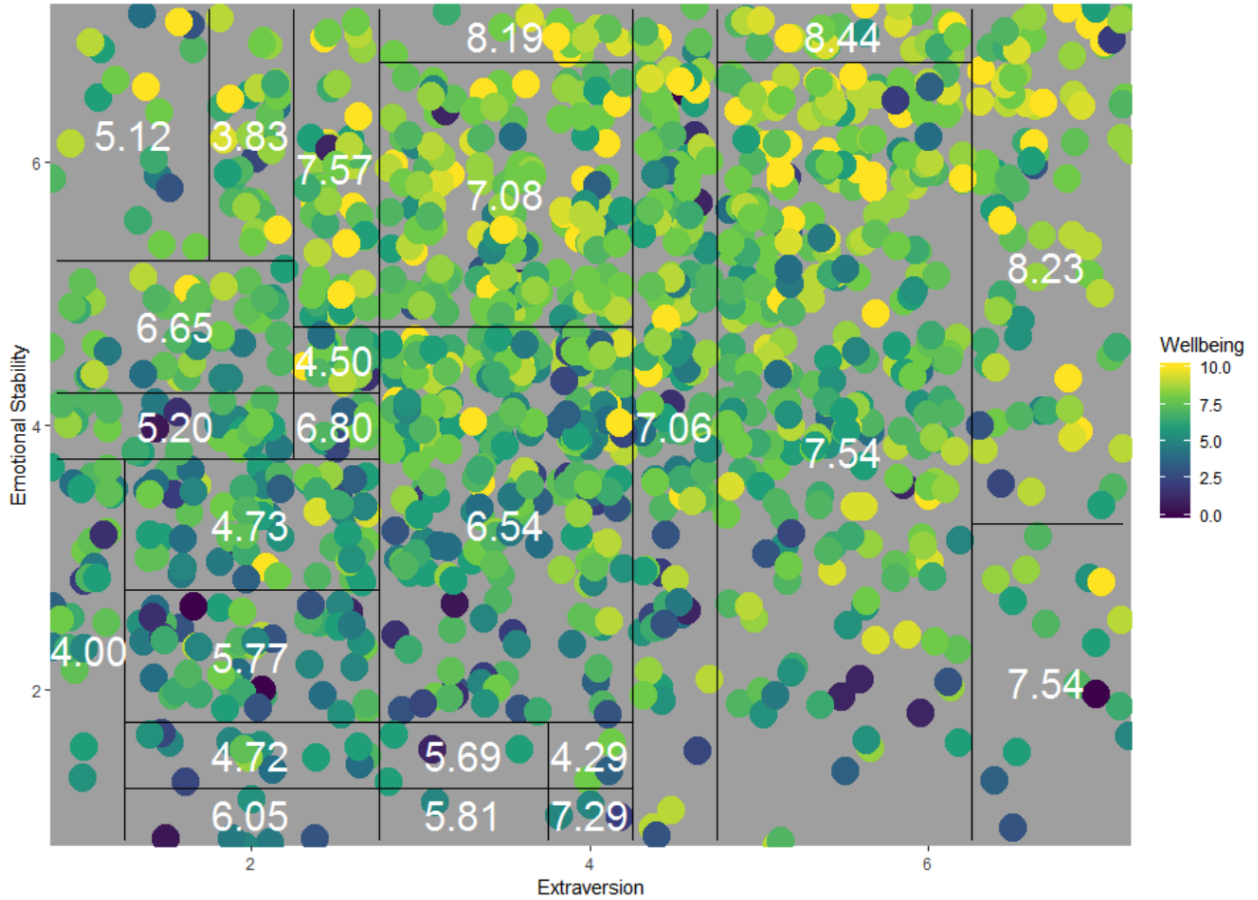


Figure 2.2: How a features space is partitioned using a decision tree algorithm. Note how the lines to split the data are always perpendicular.

The decision tree algorithm chooses where to split the data by iterating through all the input variables, and points on the variables, to find the split-point which gives the greatest decrease in variance. This can make it an inefficient algorithm if there are a lot of independent variables to choose from. The chosen point is the split which minimises the sum of the within-group sum of squares (SS) of the two new subsamples. The splitting criterion is $SSP - (SSL + SSR)$, where SSP is the SS for the parent node (see Equation 2.1), and SSR and SSL are the SS for the right and left nodes respectively. The greater the value, the better the split. This is the equivalent to maximising the between-groups sums of squares (SS) in an ANOVA.

$$SSP = \sum (y_i - \bar{y})^2 \quad (2.1)$$

The decision tree algorithm is memory-less (there is no backwards processing), so each subsample and each split decision is considered in isolation. Thus, the same feature point can be used multiple times at different places in the tree structure. This is also why it doesn't matter for the prediction if variables are highly correlated, as each variable is used independently to split each subgroup. Unless specified otherwise, the tree will continue to split until the optimal solution for the data has been reached. To prevent the model being too specialised to the training set (overfitting), the tree is pruned back so that the end 'leaf' nodes retain some variance. The optimal point to stop pruning is where the cross validation error (a measure of the tree's generalisability to new data) is lowest. This is found using the

Predicted Residual Error of Sum of Squares (PRESS) statistic (Tarpey, 2000), calculated as the sums of squares of the residuals from a sample of predicted (by the model) and observed (actual) values. If the algorithm was left to run without being pruned back, it would stop when it could predict with 100% accuracy (i.e. each datapoint was in its own individual leaf node). This solution would not generalise well to new data, and hence is why pruning or some additional stopping criterion is required.

Decision trees are advantageous as they make no assumptions on whether the relationships are linear or not. Furthermore, non-linear relationships do not need to fit a mathematical function. Decision trees also make no assumptions on the distribution of the data, meaning no transformations (for example to give the data a Gaussian distribution) are required. Additionally, any type of variable can be entered into the same model (i.e. categorical, nominal, and continuous). Moreover, decision trees provide an interpretable output from which the important variables, and important points on the variables can be identified (see Figure 2.1).

Decision trees also allow us to quantify how useful the different variables are in making predictions through the calculation of variable importance. Each split's effectiveness is determined by how much that split decreases node impurity using the SS, MSE, gini index, entropy, or information gain. To assess the performance of a variable over the whole tree, incremental node impurity is calculated. This is the sum of the decrease in node impurity for each split that the variable is used as a primary splitter (used in the tree) or a surrogate variable (a 'back-up' splitter)¹. The surrogates' contribution to variable importance is usually weighted by the strength of the agreement with the primary splitter². Different metrics and methods of calculating variable importance, including using permutations after the model has been built, can yield slightly variant results.

Decisions Trees and Non-linearity's

It is important not to assume that the structure of the data is linear, the same as it is important not to assume that if variable A effects person X in one way, it also effects person Z in the same way. Detecting and understanding sub-populations within a sample could alter the methodological approach taken or the model specifications used. It can also change how a model is interpreted. For example, Simpson (1951) noted that a statistical relationship in a population could be reversed within subgroups of a population (see Figure 1.2). For instance, "It may be universally true that drinking coffee increases one's level of neuroticism; then it may still be the case that people who drink more coffee are less neurotic" (Borsboom et al., 2009). This paradox can arise when conclusions made from one level (i.e. the individual) to another level (the population), or from cross-sectional data to intra-individual change over time (see Kievit et al. (2013), for further illustrations).

¹The surrogate splitter is not the next best splitter (which might have a different solution to the primary splitter) but instead is the variable and split point which best imitates the solution of the primary splitter.

²Note: this can be turned off in R using the parameter *maxcomplete*.

It is well documented that many factors (age, gender, SES status, culture, personality, upbringing, previous experience, and even genetics) can heavily influence how people interact with the world. Thus analysing the effect of variables across different sub groups is important. This is particularly the case in big data, where aggregating over large groups can provide a misleading account of individual processes (Wagner, 1982). One way to test whether a sample has subgroups in relation to the dependent variable, or substantial heterogeneity, is to compare the performance of a model like a decision tree (or random forest), to a simple linear model (i.e. regression). If the decision tree performs significantly better than the simple linear model, it is likely because there are sub-groups or other non-linearities present in the data, which the tree is able to capture.

Let us take the relationship between well-being and emotional stability in Figure 2.3, which is well documented in the literature to be linear (e.g. Wenzel et al., 2015). There is a significant linear positive relationship between these variables ($\beta=0.52$, $F(1,1346)=246$, $p<0.001$, $R^2=0.15$). Nonetheless, a significant relationship does not mean that a linear model predicts best. Moreover, relationships are rarely perfectly linear. Figure 2.4 shows how a decision tree carves up the same feature space when using when, rather than being constrained to be linear, it is constrained to drawing perpendicular lines. Note that the non-linearity uncovered by the decision tree does not fit a particular non-linear function (i.e. exponential, logarithmic, power, Gaussian, or Lorenz), and thus the type of non-linearity does not need to be specified, instead it learns the non-linearity from the data.

By comparing figures 2.3 and 2.4 it can be viewed how different models can find different relationships in the data. In this case, the decision tree has an R^2 of 0.16 on the sample data and a cross validated R^2 of 0.1. (The distinctions between these two measures of variance explained is important and will be explored further in Section 5.2.1.) These measures of variance explained by the decision tree suggest that using six binary splits is at least as good as, or better than the linear model, at explaining the variance in the data-set. Figure 2.5 illustrates that with more data (386,374 compared with the 1,346 used in Figure 2.3 and 2.4) non-linear relationships, which were not as pronounced in smaller samples, can be viewed and analysed.

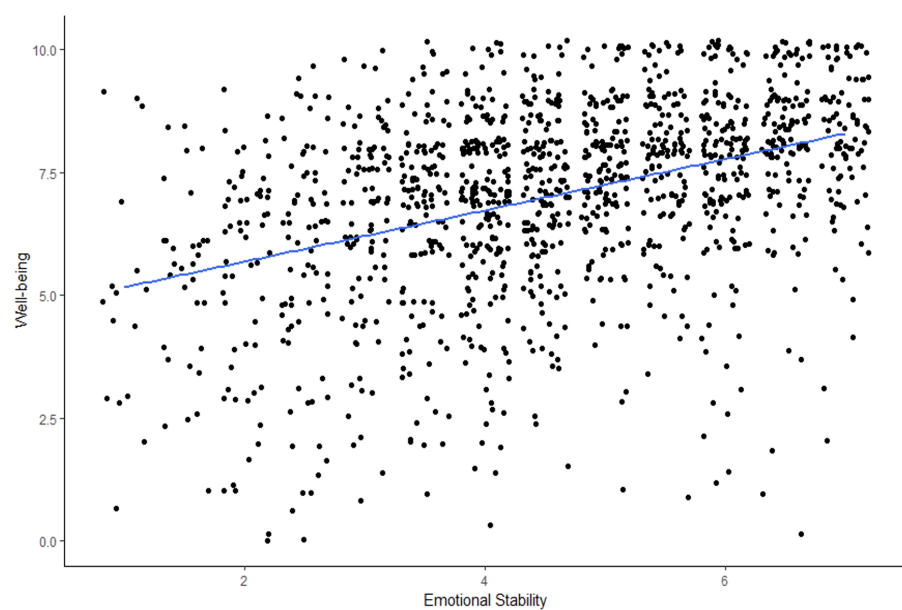


Figure 2.3: The relationship between well-being and emotional stability with a linear regression line fitted

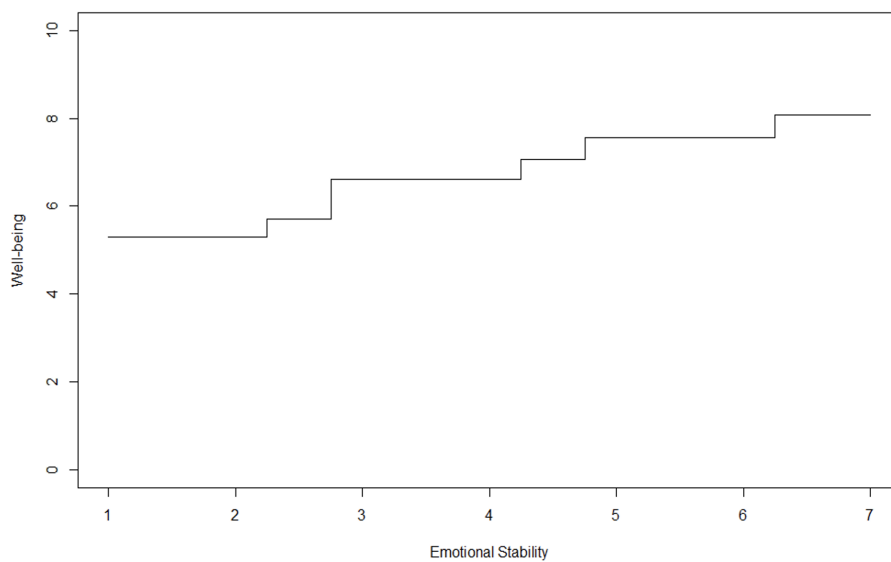


Figure 2.4: The relationship between well-being and emotional stability as decided by a decision tree

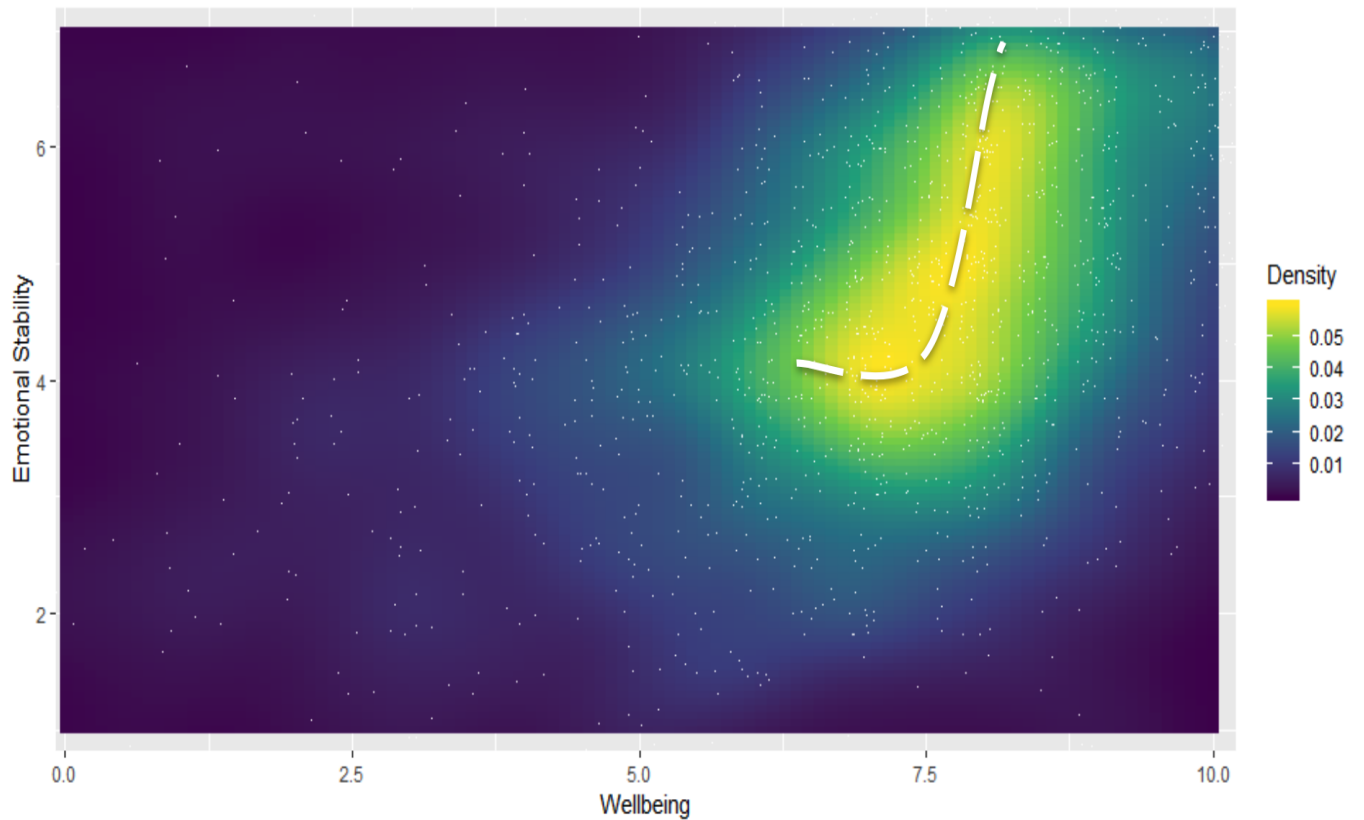


Figure 2.5: The relationship between well-being and emotional stability with large amounts of data ($N= 386,374$) and density colouring applied to aid visualisation

Limitations

One drawback of working with decision trees is that they can be unstable models. Small changes in the data can lead to large structural changes in how the tree is constructed, which can then impact the variables importance ratings. If a small change occurs in the upper branches of the tree this can have cascading effects lower down. This entails that trees built on different sub-samples of the data are often variable (Strobl, Malley, and Tutz, 2009). When using a tree to predict, the instability problem is mitigated by K-fold cross-validation procedures (detailed below in Section 2.1.4) by taking the mean error across the number of folds (K). When using decision trees to better understand the data structure or interaction effects using the tree structure, more information is better, and with more information the tree is more stable, thus fitting the tree to 100% of the data is the preferable option. A solution to the instability of single trees is to use them as part of an ensemble, for example in a random forest algorithm. Random forests often improve the prediction, as they generalise better to new data (Breiman, 2001). However this is at the expense of a granularity-loss, as the decision tree structure can no longer be viewed.

2.1.3 Random Forests

This thesis utilises random forest algorithms as they provide more stable estimates of variable importance rankings, and more accurate predictions, compared with decision trees. By producing many uncorrelated decision trees, random forests prevents the model from being sculpted to one dataset (overfitting), and therefore reflects more generalized relationships.

The prediction produced is an average across all of the trees, and because of the law of large numbers, is usually more accurate on out-of-sample predictions than a single decision tree. Random forests not only use samples of the data, they sample from the pool of possible predictors as well. This increases the variance between the trees. The common practice is for each tree within a random forest to have a random sample of 30% of the possible predictor variables Breiman (2001). Random forests are more stable as they smooth the hard decision boundaries that occur in trees through averaging. This is demonstrated in Figure 2.6.

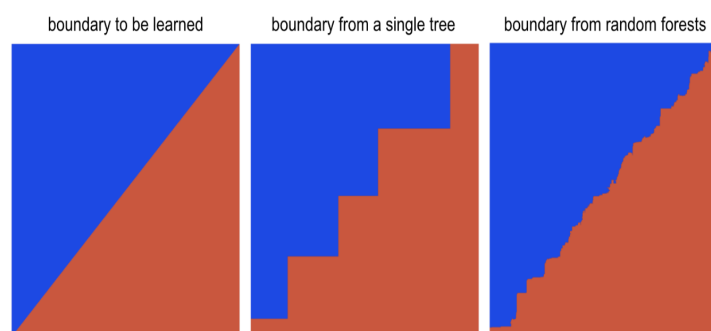


Figure 2.6: An illustration of decision boundaries being smoothed by random forests, as they take an average of many trees with high variance between them. Graphic courtesy of Deng (2018).

The negative consequence from improved prediction accuracy and model stability through the use of random forests, is a reduction in interpretability - as the tree structure cannot be displayed. Due to each local decision depending on the last local decision, and the different trees being forced to make a diverse set of decisions due to the different samples of data and predictors, there exists no ‘average tree’ (Strobl, Malley, and Tutz, 2009). However, variable importance measures can still be calculated, on an individual tree basis and then averaged over the forest. This makes the variable importance measures for a random forest more stable than for an individual tree.

2.1.4 Cross Validation

Cross validation is used to recycle data rather than ‘throwing it out’ in the training phrase. Cross validation is a technique which is used to validate the models’ performance on novel data. It is seen as the next best option to a full replication on new data (Yarkoni and Westfall, 2017). The data is first split into a number (K) of sections or folds. If $K=10$, for example, the model is built using 9 of these folds, and then tested on the remaining one. This process is iterated 10 times, so each of the 10 folds has been used as the test fold. The mean of the error for the 10 test folds is taken as the measure of the model’s performance.

Cross validation is also often used to optimise models on the training data. This process is called the validation phase. This can involve using K -fold cross validation to try out different meta-parameters to find the optimal combination. This is commonly done using a grid search algorithm which repetitively evaluates a range of different possible parameter values, and selects the meta-parameters which produce the most accurate model. After a

validation phrase, researchers can be confident that they have the optimal model for them to test on new data. The validation phase can be conceptualised as optimising the model to generalise on new data, and thus preventing overfitting. The optimised model’s performance is then evaluated on new data to explicitly *test* how well it generalises.

Set out in the Introduction is the proposal that big data has the potential to augment psychological research. Yet the scientific approaches used to research in big data are often at odds with those used in psychology. This led to the current piece of work merging theory driven and data driven approaches into a loosely defined sub-field *big data psychology*. Big data psychology is proposed here as its own domain of psychological research, which straddles many topic areas, due to the fundamentally different approach and methodology required (compared to that currently being used by the majority in psychology).

2.2 What would Big Data Psychology look like?

The previous section in this chapter outlined the different machine learning methods which are afforded by big data, yet how do these methods have the potential to enhance psychological research, as proposed in the Introduction? This section outlines the area named in this thesis as “big data psychology”, which emerges when the discipline of psychology meets machine learning methods. It is henceforth proposed how the two fields could come together to enhance our understanding of human behaviour in the real world.

Psychology is predominantly a theory driven discipline. A theory is an idea, or set of principles, that can provide an explanation for phenomena. In psychology, a theory is used to provide an understanding of the world for humans; whilst a model is used to predict phenomena, operationalising theory³. Theories allow us to interact with the world by simplifying it, reducing the complexities of the world into a set of rules which are usable. For example, in physics, the same laws of thermodynamics can be used to calculate the properties (such as temperature, pressure, density) of any substance in equilibrium. In many ways, the critical standpoint that has developed so extensively in the scientific research of instrumentalism, is yet to fully transfer to psychology. The global instrumentalist holds that theories (similarly ideas and concepts) are tools which help us to navigate through the world. This requires that theories are *useful*, but does not compete with their ‘truthfulness’, as a useful theory can be ‘wrong’. This is also the approach in machine learning which focuses on the utility of a model through prediction accuracy. Yet, psychology often forgoes prediction accuracy for simplicity and theoretical understanding. Thus the goal of big data psychology, where machine learning meets psychological theory, could be described as using big data to build useful prediction models, yet also exploiting the vast amounts of information to improve our theoretical understanding of the behaviours being predicted.

³A model and a theory across many disciplines are used interchangeably. Yet here, in line with the common linguistic practice in psychology, a model will refer to the instrument used to test *specific* hypotheses (which are often derived from a broader theory).

Why do we Need Explanations?

It is much harder to test if an explanation is correct, compared to a prediction. Predictions allow us to assess the usability of models, so why then is prediction alone not enough? Here, this research argues there are three reasons for why explanations are important. Firstly, if we want to design or undergo an intervention, we need to understand the causal pathways. This involves both a prediction and an explanation. We want to be able to predict the effectiveness of a given intervention, and we want to understand *where* to intervene. Thus, understanding the important variables in a model is crucial for efficiently targeting resources and to maximise real world impact. Secondly, we also want to know *why* a model did/did not work. If we understand this, further improvements/developments to the model and/or the intervention can be made. In this way, explanation can improve prediction. Finally, it cannot be ignored that humans are inherently curious and it is satisfying for us to ask the question *why*, and we naturally do this, even if there is no use. Therefore, this research proposes the use of both data derived methods for predictions and theoretical interpretations within the domain of big data psychology.

Deriving Theories from Big Data

Traditionally, deduction is used to produce theory. Models are constructed to test theories by way of hypotheses, with the hope that they will prove to be predictive. However, if the theory driven model isn't useful (it can't produce good predictions), then the process starts again, and another theory needs to be 'thought up' and tested. This approach to knowledge discovery can be slow, and it implies that a theoretical reasoning is sufficient to correctly specify the optimal model parameters. Model specification is crucial. If it is not done right a model can produce poor or wrong predictions, even if the theory is correct. Moreover, a theory could be sound, but not specific enough to provide accurate model specification. If it is simply assumed that theory produces the best model specification, and this assumption is not tested, important ways to improve the model, and consequently enhance the theory, could be lost.

Alternatively, theory can be derived from models that have already been shown to be useful in the world. We can find the best model, which we know produces useful predictions, and deduce theory from them. In this way, finding the optimal model to represent the world ultimately comes down to trial and error. Traditionally this has been too computationally costly. However, with today's computing power and optimisation algorithms, this is now a viable option. This, and the availability of large amounts of data, allows us to model the world using a bottom-up inductive approach. The resulting models are built upon what happens in the real-world (rather than theory). If theory can be derived from such models then it has a greater chance of being practically useful.

A Big Data Psychology Approach

$$\text{Deductive Science: Theory} \rightarrow \text{Model} \rightarrow \text{World} \quad (2.2)$$

$$\text{Inductive Science: World} \rightarrow \text{Model} \rightarrow \text{Theory} \quad (2.3)$$

$$\text{Big Data Psychology: World} \longrightarrow \text{Model} \longrightarrow \text{Theory}$$

$$1. \text{ World} \longrightarrow \text{Model} \quad (2.4)$$

$$2. \text{ Model} \longrightarrow \text{Theory}$$

Whilst predominantly deductive sciences like psychology use theory to build models (which can predict things in the world) (see Diagram 2.2), inductive sciences like the field of machine learning often use real world data to build and tune models (which can lead to theory) (see Diagram 2.3). Psychology is good at Theory $\rightarrow$ Model, but often falls short on the generalisability of the model to the real world (Open Science Collaboration, 2015). Machine learning does well at using complex algorithms to model the world (World $\rightarrow$ Model), but falls short on using these models to generate theory on how the data was generated. Big data psychology has the potential to take the best from both disciplines and try to generate useful theories of human behaviour that generalise to the real world. This thesis proposes that by using machine learning on big behavioural datasets, accurate representations of the real world can be encoded in a model for prediction. From that model, insights can be extracted which can feed into the psychological theory of behaviour. This is conceptualised as a two step process (see Diagram 2.4) involving 1) creating a prediction model of the world, and 2) trying to interpret and explain the model. This dual approach utilises the high predictive accuracy and applicability of modelling behaviour in large datasets; and combines it with psychology's methods for explaining and interpreting models for theory. Such an approach requires a broad remit of models, methods, and approaches, and is the sphere, hereby referred to as big data psychology, that this thesis operates in.

2.3 A Review of Big Data Psychology Research

Big data psychology can cover many topics, as referred to above, and thus what qualifies as big data psychological research is not well defined. As well as the ambiguity around what qualifies as big data (as outlined in Chapter 1), psychology also has similarities and overlaps with many other disciplines. For example, both the fields of health research and psychology investigate mental health, food choice, and sleep. Similarly, as the movement towards utilising big data is occurring in psychology, equivalent movements are occurring in other related research disciplines. Thus, before focusing on the research strictly within the discipline of psychology, this review first describes the wider context within which the movement towards big data psychology occurs.

The literature review starts broad, outlining first the other areas in which big data has been used as a measure of behaviour in research; it then identifies the ways in which big data has been used in research related to psychology in the peripheral fields to the work in this thesis - health and the environment. Within the psychology literature big data has mostly been addressed for consumer and marketing purposes or as a demonstration of big data methodology to psychologists. The literature review will delve into both of these primary research areas before concluding with a critical review of the smaller body of work which is most related to this big data psychology thesis.

It is proposed here that for research to constitute as big data *psychology* what it should aim to address is three-fold. 1) It should look to provide greater insights or *understanding* into the possible drivers of behaviour (or psychological phenomena) - prediction alone is not sufficient. 2) It should attempt, where possible, to study *individuals* (Lamiell, 2015), and finally 3) it should look to contribute findings which are relevant or useful to psychological theory.

2.3.1 Data as a Behavioural Measure

In recent years, big data has been used as a behavioural measure for researchers studying human behaviour over a number of different domains. For example, travel behaviours using smart cards (Chu, 2015), tourism behaviours using social media (Miah et al., 2017), consumer behaviour using transactions (Abdulla et al., 2015), crowd behaviour using open content web sources (Kallus, 2014), commercial competition using online search data (Ringel and Skiera, 2014), online traffic using web activity (Yoon et al., 2014), and in different ways in the health domain using data collected from applications (Hermon and Williams, 2014). These studies show that a variety of different big data types are now being collected and made available for academic research into human behaviour.

Yet, although big data research is being used to measure behaviour across a number of different domains, such behaviour is predominately aggregated and de-personalised. Thus, such datasets are used to study trends of people at mass, and although they can be used to derive something about human behaviour in general, will likely not be utilised to study individual differences in human behaviour. What makes purchasing data from loyalty cards so useful, is the ability to tie all purchases (when the card is used) back to one unique identifier, so a person's behaviour can be tracked over time. Further, deriving something about an individual's psychology is hard unless ground truth information is attained. The relationship between two behaviours, which may be proxies for some kind of psychological state (such as the purchasing of health food, and anxiety medication) can be analysed; yet exactly what these proxies are for, and what they are measuring, is hard to delineate without ground truth information.

2.3.2 Related/ Peripheral Fields' use of Big Data

The relatively new trend towards the uptake of big data use in psychology is not occurring in isolation. There are movements in this direction in psychology's peripheral fields as well. This thesis focuses primarily on mental health and environmental behaviour measured through transaction data, thus the literature review focuses on these domains. Although there are few studies that use big data in the domain of environmental psychology or the psychological study of mental health, research in health/medicine journals and environmental science journals, in which big data has been exploited, has been conducted in parallel to psychology. This research is outlined in brief below.

O'Connor (2014) writes that (similar to psychology) the healthcare industry in general has been slow to embrace Big Data. This has been attributed to the cost of adding analytic functions to existing electronic health records, privacy issues, poor-quality data, and a lack of willingness to share data. Yet, in health journals, studies are using big data in innovative ways to study psychological phenomena such as mental health. For example, Ayers et al. (2013) assessed the viability of aggregate internet search queries for real-time monitoring of several mental health problems, specifically in regard to seasonal patterns in the searching for mental health information. Dzogang et al. (2017) similarly studied seasonal affective disorder (SAD) in digital traces such as the language used in tweets from Twitter, and the purchasing of over the counter anxiety medication. As well as using secondary big data, other researchers are using applications on peoples phones to actively collect data. For example, Harrison et al. (2011) described an initial proof of concept study assessing the potential utility and effectiveness of a newly developed mobile phone and web-based program MyCompass to help manage mild-to-moderate stress, anxiety and depression. Thus, big data is being utilised for studies in mental health research, of which many have applications in psychology. Yet, much of this work is predominately being carried out by health researchers and being published in healthcare or clinical mental health journals.

In the field of environmental psychology, until recently, environmental behaviour has been measured using surveys, by observing people in the real world without them knowing (Poortinga et al., 2016; Kormos and Gifford, 2014), gaining peer ratings (Chao and Lam, 2011), or getting people to report household meter readings (Fuj et al., 1985). Recent research has started to use data from smart meters to study environmental behaviours (Kavousian et al., 2013; Zhou and Yang, 2016). For example, Sønderlund et al. (2014) described the effectiveness of water consumption feedback technology in promoting water saving behaviour, and De Dominicis et al. (2019) showed that making a smart meter take on a social component can promote long term energy consumption. Yet the problem with such smart-meter data, is that it is rarely at the individual level, as most people live with other occupants. Thus, such studies are limited by the extent to which the findings are attributable to the individual characteristics of the participant in the study.

2.3.3 Marketing Sciences and Consumer Psychology

Within the discipline of psychology, the fields where big data use has been greatest is consumer and marketing psychology. Applications of Big Data in these fields have largely focused on assessing customers' preferences (Jacobs et al., 2016), predicting what customers are most likely to buy next (Ghose et al., 2012; Lu et al., 2016; Linden et al., 2003), improving targeted advertising (Hauser et al., 2009), understanding brand perceptions (Culotta and Cutler, 2016), and describing the markets (Netzer et al., 2012) (see Wedel and Kannan (2016) for a review). Because such research has strong practical commercial applications the research has been driven by prediction, rather than trying to *understand* consumer attitudes, emotions, and behaviour (Matz and Netzer, 2017). Davenport et al. (2001) highlight this, stating that although holding vast amounts of information on customers can help businesses to 'know more about their customers' it won't necessarily allow them to 'know the customers themselves' (referring to their psychology).

For Marketeers, personality data is a valuable tool to have. Personality has been shown to predict consumers behaviour (Tkalcic et al., 2009), and unlike affect, emotions, interests, and attitudes, personality traits are theorised to be stable over time (Costa Jr and McCrae, 2008; John et al., 1999). This entails that just a single measure in time is required to help predict an individual's consumption behaviour for life. For example, recommender systems using collaborative filtering algorithms can make predictions about many items a person will consume using a personality test, opposed to multiple ratings of how individuals interacted/reacted to similar items in the past (Tkalcic et al., 2009).

To attain information on customer's personality, researchers have to survey many individuals, making such data difficult to acquire. To try to circumvent these measurement issues, researchers have tried to infer personality from consumer behaviour (Gladstone et al., 2019). If done accurately, this negates the need for personality data collection on a large scale, and produces insights about the types of behaviours related to personality types. Researchers have demonstrated the ability to infer personality traits from personal websites (Marcus et al., 2006), Facebook or Twitter profiles (Quercia et al., 2011; Kosinski et al., 2013), language use (Park et al., 2015; Schwartz et al., 2013; Pennebaker and King, 1999), and more recently shopping behaviour (Gladstone et al., 2019).

Most of the research on personality and consumer behaviour has focused on The Big Five personality traits (see Chapter 3), measurable in as little as 10 questions (Gosling et al., 2003). Yet there are many other ways that individual differences can be conceptualised and measured - for example using time preferences, impulsivity, frugality, shopping impulsivity, self control, and well-being (all of which will be discussed in relation to consumer behaviour in Chapter 3). In summary, the big five have been exploited in big data research in consumer psychology, but primarily with the goal of enhancing prediction models or for commercial benefit, and less for informing psychological theory.

2.3.4 Literature Informing Psychological Theory

Recent research demonstrates that big data can be successfully employed to study important psychological phenomena ranging from personality (Youyou et al., 2015), language (Schwartz et al., 2013), emotions (Kramer et al., 2014), cultural fit (Danescu-Niculescu-Mizil et al., 2013), social networking (Ugander et al., 2011), as well as shopping habits and well-being (Matz et al., 2016). Although these examples may inform psychological theory, the lead author of Matz et al. (2016) (who's study is reviewed in more detail in Chapter 4) is the only researcher (focusing on the first authors) who has a psychology background. The other authors include an applied mathematician situated in an engineering department, a researcher in information systems, a computer scientist, a researcher in management, and a data scientist at Facebook. Furthermore, the journals in which these pieces of work were published in (with the exception of Matz et al. (2016)) were all multidisciplinary journals, or computer science conferences such as world wide web (WWW). Although multidisciplinary research can undoubtedly be a great thing, leading to unique and interesting work, the lack of psychologists' involvement in such work, and research being published in non-psychology locations, can distant psychologists from big data psychology research.

2.3.5 Big Data Methods and Tutorials

Part of the movement towards big data being utilised in psychology (and other disciplines outside of computer science) involves a push for the potential skills gap (as outlined in the Introduction) to be overcome. This potential gap has three main components: 1) data acquisition/storage, 2) data analysis/ model implementation, and 3) model evaluation. Articles have been published in psychology methodological journals on the techniques which can be employed for research in big data. These span all three of the aforementioned components from data acquisition to model evaluation.

Much of the focus is on using unstructured social media data from Twitter or other online sources, as these are an easy and accessible way for anyone to acquire big data. Landers et al. (2016) describes methods on how data can be acquired from open source websites using a web-scraper. They highlight the need to emphasise theory when dealing with messy unstructured data, in order to collate it in a way which can produce relevant and meaningful insights. A methodology is proposed to mitigate against the possibility that the findings are just happenstances of a large collection of information. Once aggregated, this data can be used to analyse societal trends and norms in social consciousness. For example, Jones et al. (2016) used twitter data to study the change in language use after traumatic events. Similarly Stanley and Byrne (2016) used social media to study computational models of long-term memory.

Social media offers a great opportunity for psychologists interested in social interactions to acquire large datasets. Yet, having the technical capabilities to extract the data is not enough. Once unstructured text data is collated, methods which psychology researchers

may not be acquainted with (such as machine learning models and cross validation), are required to organise and analyse it. Kosinski et al. (2016) and Kern et al. (2016) outline methods for analysing social media data such as dimensionality reduction (using single value decomposition (SVD)) and generative probabilistic models (such as latent dirichlet allocation (LDA)). Other popular analytical methods include latent semantic analysis, topic modelling, bag of words approaches, stop words, and support vector machines to analyse text data (Chen and Wojcik, 2016).

As outlined in the Methods section, the research in this thesis focuses particularly on decision trees and random forests. Decision trees have been used to model behaviours such as adolescent alcohol consumption (Gervilla et al., 2010), cannabis and cocaine use (Gervilla and Palmer, 2010), to segment consumer behaviours (Abdulla et al., 2015), and to identify the variables most predictive of suicide attempts (Mann et al., 2008). Random Forest models, an ensemble of decision trees (see Method section), are also being exploited in psychology for a growing number of applications. For example, Oh et al. (2003) and Shen et al. (2007) applied random forests to neural ensemble recordings and EEG data (high dimensional data, which standard regression models would fail with). Random Forests have also been used on lower dimensional data in psychology on determinants of once-only contact in community mental health service (Rossi et al., 2005), and predicting attempted suicide using family history (Baca-Garcia et al., 2007). However, in comparison to other methods in psychology, such as regression, decision trees and random forests are not as well known as an analysis method to psychologists, as they fall outside the standard statistical syllabus.

Working with big data also requires special methods for performing model evaluation such as cross validation or out-of-sample testing. Such methods have been demonstrated for a psychology audience in methodological papers. For example Kern et al. (2016) explained how to implement the k-fold cross-validation. The method is outlined clearly, but what is lacking is discussion and tuition on *why* these methods are being employed, and what are the consequences if small data evaluation methods and statistics are applied on large datasets. If researchers are not grasping the implications of over-fitting, for example, they will be less conducive to implementing methods to mitigate it.

Further methodological articles are outlined in a special Big Data edition of Psychological Methods (Harlow and Oswald, 2016), and explained succinctly by (Oswald and Putka, 2015) for organisational psychologists. In many cases these tutorials are accompanied by code and test datasets. This wave of methodological papers for psychologists is a crucial step in addressing the potential skills gap. However, simply having the tools to learn something, does not mean that these resources will be utilised. As outlined in the Introduction, not only may there be technical barriers to big data psychology, but perhaps more important is addressing the cultural differences which occur between machine learning and psychology.

Many of the big data methodological papers in psychology lack a conceptual discussion on *why* methods such as cross validation, are vital for big data research. *Why* multivariate linear regressions should not be fitted to big data. And *why*, analysing big data may require

an inductive rather than deductive approach. Furthermore, the nuances of working with big data are not discussed or explained. Reading such tutorials, readers would be forgiven if, once the data was collated, they expect it to behave the same as small data. The intricacies of data cleaning, data validity, noisy data, sparse data, effect sizes, and the utility (or not) of p -values are rarely fully incorporated into the methodological papers for psychologists using big data.

2.3.6 Summary of the Gaps in The Literature

This review of big data psychology research highlights that many big data method and tutorial papers are making their way into psychology journals, perhaps as an attempt to reduce the potential skills gap and encourage more psychology researchers to take up big data research. The application for such methods in psychology, however, is predominately in the fields of consumer psychology and marketing sciences. In such research, psychological phenomena, such as personality traits measured using the big five, are being utilised to predict behaviour in digital traces. Yet, less research is focused on using such data to better understand the psychological processes themselves (Bleidorn et al., 2017). This trend is reflected in most big data studies which cross over into psychology. To summarise, three main gaps have been identified in the literature which uses big data for psychological research. These, as well as how this thesis addresses the gaps, are as follows:

1) A paucity of big data psychology research is informing psychological theory (outside of the marketing/consumer domains). Big data psychology is dominated by marketing science and consumer psychology, and rarely have the potential benefits of big data been realised in health or environmental psychology. Further, when marketing science and consumer psychology research have incorporated psychological theories, such as personality, into their models it is to better *predict* behaviour in big data. Little research has been conducted to interpret and understand the predictive relationships; and less still has used predictive models to evaluate the conceptual structures of personality (Bleidorn et al., 2017). Research in other domains, running parallel to psychology, have been more involved in harnessing big data to measure human behaviour. Although such findings are often applicable to psychology, much of it is not reported in psychology journals, and not conducted primarily by psychologists. This can mean that psychologists don't engage with the big data psychology literature as much as they could, perhaps exacerbated by the different methodologies being used.

This thesis brings big data psychology to psychologists, particularly the fields of health and environmental psychology where big data is utilised less. This thesis also aims to demonstrate that big data and personality questionnaires can be combined to better understand the psychological drivers of particular behaviours, not just to improve marketers' prediction models. And finally, this research aims to show that transaction data can be used to understand a whole range of different behaviours, not

just those directly linked to consumption and consumer psychology.

2) A lack of engagement in when and why big/small data approaches and analyses are appropriate. Methods journals in psychology have published special editions on the methods of big data, including tutorials and examples of how psychologists can extract and process big data to analyse it in a statistical way. But rarely is it discussed why methods (such as cross-validation) are useful, or the fundamental ways in which small data and big data vary. In addition, there is minimal engagement with the weigh-up between prediction and explanation, or the utility of an inductive versus deductive approach.

This thesis critically engages with the different scientific approaches within big data psychology. It also suggests frameworks under which both prediction and explanation can be conducted with the goal of extracting maximally accurate insights from data. It suggests ways in which both induction and deduction can be combined together, negating the need to reside in just one camp. Crucially, it addresses why different approaches are required for big data, and the issues that arise when using small data statistical approaches on large datasets.

3) A lack of investigation into the psychological drivers of behaviour on an individual-level. Researchers, across multiple domains, have harnessed big data as a way to study aggregate behaviour in the real world. However, because of issues with privacy and data availability, few have used big data to study the individual level drivers of behaviour. It requires a unique dataset and opportunity to enable the linking of individual level survey data and digital data. A study by Matz et al. (2016) (which will be discussed in depth in Chapter 4) is a rare example of where this has been done using transaction data. (Yet this was for a much smaller sample ($N=625$) than that used in this research ($N=12,968$)).

This thesis exploits the unique opportunity afforded by Boots to link real world transaction data with individual's responses to a questionnaire measuring personality and other psychological phenomena. This resource underpins the big data psychology research in this thesis, providing an exciting opportunity to gain ground truth information on the individual's intrinsic states and traits, and to study the relationship of these to individuals behaviour in mass transaction data, as it happens in the real world. These two data sources (the questionnaire and the transaction data) are outlined in detail in the following chapter.

Chapter 3

Methodology and Data

It was highlighted in the previous chapter, that one of the unique aspects of this research, which enables psychological investigation using big data, is the opportunity to connect transaction records (as a measure of real world behaviour) with data collected about the individuals' psychological and socio-demographic information. This allows not just the analysis of trends or correlations between mass aggregated behaviours, but the explicit use of big data as a measure of an individual's behaviour and the relationship to their self-reported personality profiles.

The methodology of this thesis utilises two main sources of data: a mass survey, and loyalty card transaction data. The transaction data (which will be discussed in more detail in the following section) is large, spanning 4 years, and containing almost 3 million customers. It is in a structured relational database, making it easier to manipulate, despite it not being collected for the purpose of research. The transaction data is noisy (holds additional meaningless information) and can contain errors, requiring data cleaning and preprocessing steps. These characteristics, among others, define the transaction data as 'big data' (Cuartas Arias, 2017). On the other hand, the mass survey (distributed by Boots to 80,000 loyalty card customer) retrieved just over 15,000 responses. It was collected for the sole purpose of research, and despite the large sample size, for this reason could be conceptualised (under some definitions) as a large sample of 'small data' (Cuartas Arias, 2017). Control over the design of the questionnaire was maintained, and the survey was piloted before it was distributed *en masse*, ensuring the information obtained was less noisy, and more reliable.

The survey required the individuals to input their loyalty card number into a text box, so the survey responses could be matched to the transaction data. The goal of linking the questionnaire with the loyalty card data was to be able to study the psychological drivers of human behaviour in a consumer context. As consumer behaviour was measured using loyalty card transaction data, no additional scoping or data collection (beyond acquiring the data from Boots) was required. For the survey, however, it had to be carefully planned what information was most desired on individuals, and how best to ask it and measure it in a survey format. Hence, before constructing the questionnaire a literature review was conducted on the psychological and demographic variables which may influence individuals' consumer behaviour. This examination of the literature is outlined in the remainder of this

section.

This chapter first reviews the factors which have previously been found to affect consumer behaviour in the literature in Section 3.1. It starts by describing the ways in which demographic variables have been shown to influence shopping, then generalised personality structures such as the Big Five and impulsivity, before looking at individual differences that directly affect how people shop, such as shopping impulsivity and frugality. The second half of the chapter, in Section 3.2, introduces and describes the data from 1) the Boots loyalty card customers transactions; 2) the survey responses; and 3) the sample of transaction data which emerged when the fully completed survey responses were linked to the loyalty card data.

3.1 A Review of the Factors Affecting Consumer Behaviour

Throughout this thesis, consumer behaviour is used to study the predictors of real world health and environmental behaviour. It is investigated whether transaction data, left behind in people’s digital footprints, can inform us about human behaviour. Specific consumer behaviours, such as the purchasing of plastic bags or over-the-counter anti-anxiety medication, may be able to provide a proxy for more general behaviour, attitudes, beliefs, and mental states. For example, if a customer purchases environmentally friendly products, this person is performing an environmentally friendly consumer behaviour. An inference may be made that this individual holds pro-environmental attitudes, and may engage in other ecological behaviours. This inference, however, may not hold. It can not be ignored that these behaviours are inherently linked to the consumer context they are occurring in, and there are a multitude of consumerist motivations that could drive a particular purchase.

A person’s mood, their bank balance, the product’s price, offers, stock levels, product quality, brand perceptions, brand loyalty, habit, the weather, fashion/trends, whether a person plans their purchases or prefers to buy spontaneously, buying online/offline, whether more pleasure is derived in saving and being efficient with money versus ‘splashing out’, can all influence purchase patterns. As a consequence, consumer behaviour may not be indicative of generalised attitudes and beliefs. To try and best depict the motivations and drivers behind particular consumer behaviours the mass survey attempted to measure many of these factors which can affect shopping behaviour.

A problem with trying to understand the drivers of consumer behaviour in big data, is that the same purchase behaviour can stem from entirely different, and even opposing, underlying forces. For example, a person who buys multiple packs of vitamin supplements may be particularly health conscious, may be unwell, or may not be able to resist the 2-for-1 offer at the till because they are an impulse buyer. These are three conceptually distinct reasons for the same purchasing behaviour. Thus, it is important to understand

what psychological mechanisms affecting consumer behaviour in general (i.e. impulsivity affecting the extent to which individuals plan shopping trips), as well as what might be related to specific behaviours (such as pro-environmental attitudes influencing ecological purchasing), so that models can be more accurately and intelligently interpreted.

A fuller understanding of the drivers of purchasing patterns requires, not just understanding the psychological drivers of health and environmental behaviour (outlined in the literature within the relevant chapters), but also the drivers of generalised consumer behaviour. The following section outlines how general demographics and personality variables can influence consumer behaviour, before looking at individual differences in how people shop. Because of restraints on how much time the questionnaire could take (roughly 20 minutes), to minimise the drop-out rate, not all factors influencing consumer behaviour could be included in the questionnaire. Therefore the following literature review was used to help decide on the most important questions to include in the survey. The psychological variables which were not explicitly measured are still discussed here in brief, as the contextual theory helped with model interpretation throughout the thesis. (The exact features and the scales used in the questionnaire are outlined in the data descriptions in Section 3.2.)

3.1.1 Demographic Features

Demographics have been used to try to explain differences in consumer behaviour patterns, with varying degrees of success. The literature predominately focuses on gender differences in consumer behaviour, which are outlined in detail in the next sections. Other demographic variables which have been investigated include income, social economic status (SES), and social relationships. An Indian study found that marital status, family size, and income influenced online shopping behaviour (Nagra and Gopal, 2013). Whilst Hernández et al. (2011) did not find that any socioeconomic variables influenced online shopping behaviour for “experienced” internet shoppers. Race was found to have no effect on online shopping, and marital status only a marginal effect; whilst higher education appeared to make people less successful at searching for products online (Ward and Lee, 2000). Research on the effect of income and SES on consumer behaviour has heavily focused on food consumption and health. Studies have investigated geographic measures of food availability (Walker et al., 2011), the impact of SES on nutrient intake in US households (Adrian and Daniel, 1976), and the effect of SES on food choice (O’Dea and Caputi, 2001). As well as SES, education level has also been shown to influence on consumers’ dietary and health behaviour (Petrovici and Ritson, 2006).

Those who have a higher income, are understandably likely to spend more money (Landis and Gladstone, 2017). Interestingly however, some studies suggest that those living on a low income tend to spend a higher percentage of their money on products and experiences which are perceived to hold a higher status (Bagwell and Bernheim, 1996). This ‘wealth signalling’ through consumption behaviours enables poorer groups to restore feelings of self-integrity

(Sivanathan and Pettit, 2010) and personal power (Rucker and Galinsky, 2009). This is done more by individuals who are high in extraversion (Landis and Gladstone, 2017). Thus, the relationship between wealth and consumer behaviour is not as linear as one might first expect.

Forsythe and Bailey (1996) showed that in general the factors affecting time spent shopping were “complex and interrelated”. Significant main effects were found for shopping enjoyment, marital status, and age. They additionally showed that the positive effect of shopping enjoyment on time spent shopping was mediated by marital status. Conversely, Bhatnagar et al. (2000) found that marital status had no influence on online shopping across seven categories. Bhatnagar et al. (2000) highlighted that the type of shopping matters. Age predicted online purchasing of music (negatively), books (positively), travel (positively), but no there was no effect found for films or clothing. Males were more likely to buy films, and less likely to buy books and clothes than females were online, yet no gender differences were found for music, or concerts/plays. The ways in which age and gender have been shown to influence consumer behaviour is outlined in more detail in the sections below.

Age

Age has mostly been studied in relation to impulse buying. Like other facets of impulsivity, shopping impulsivity varies with age (Han et al., 1991). Younger people feel less risk when spending (despite having less money) and are thus more inclined to buy impulsively (Parboteeah, 2005). Between the ages of 18 and 39 impulse buying has been shown to increase, and decline thereafter (Wood, 1998; Bellenger et al., 1978). Age has also been shown to affect what consumers buy, and whom they buy for. This is particularly the case for food shopping (Westenhoefer, 2005).

Gender

Well established gender effects have been found in the consumer behaviour literature. Studies show that women are greater impulsive buyers than men (Kollat and Willett, 1967; Parboteeah, 2005; Rook and Hoch, 1985; Dittmar et al., 1995; Coley and Burgess, 2003) women process advertising and product information in a more detailed and comprehensive way than men do (Meyers-Levy and Sternthal, 1991; Kempf et al., 2006; Laroche et al., 2000) women use supermarket loyalty cards more than men (Bellizzi and Bristol, 2004); women prefer to choose from an assortment of products (Kruger and Byker, 2009) and are more likely to pay attention to when items are on sale (Kruger and Byker, 2009); and women are also more likely to pick up and touch items when shopping (Tifferet and Herstein, 2012).

Kruger and Byker (2009) argue that many of these sex differences can be explained using evolutionary theory. They theorise that when our female ancestors were gatherers it was imperative that they carefully scrutinise berries, remember where they were, and when they were ripe. This theory is supported by a study which found that, at a large farmers’ market, women were better than men at identifying the location of stalls where they had previously

been given food items (New et al., 2007). In addition, Ward and Lee (2000) found that men both spend less time searching for products online and give up searching more quickly than do women do. However, it is argued that these sex differences could stem from women, in general, enjoying shopping more than men (Rook and Hoch, 1985), and thus holding a greater interest. An elevated interest in shopping could also account for why women spend more time and care picking out their purchases, particularly Christmas presents (Fischer and Arnold, 1990). Therefore, interest or enjoyment in shopping needs to be controlled for before determining where the sex difference have evolutionary origins or not.

Gender effects could be partly explained, or mediated, through personality differences found between the sexes. On average males are less risk averse (Byrnes et al., 1999) and more assertive than females (Costa Jr et al., 2001). Costa Jr et al. (2001) also found that females are higher in Neuroticism, and Agreeableness. These sex differences are not thought to be established through traditional gender roles as they are most pronounced in western cultures, where sex-role differences are smallest (Costa Jr et al., 2001).

Gender differences in consumer behaviour vary depending on what items are being bought (Bhatnagar et al., 2000). For example women purchase more frequently than men (but spend less when they do shop), and are more likely to purchase for the whole family (Kuruvilla et al., 2009). Studies on diet and food choice, often treat men and women entirely separately, as they have very distinct eating motivations and patterns. For example, Zellner et al. (2007) found an interaction between stress and gender, showing that men ate more unhealthy foods in a non-stressed condition, and woman ate more in the stressed condition. Therefore, the type of shopping and what is being purchased is important to consider when explaining sex differences. It is expected that because a large portion of the Boots data is beauty products, gender will likely have an effect on the amount and type of products people buy.

3.1.2 Personality Theory

As described above, demographics can influence shopping behaviours. Yet, variables such as age and gender are not enough to accurately segment customers, other than into broad categories. Personality, however, slices through the entire population, irrespective of demographic profiles. Thus studying the effect of personality on consumer behaviour has the potential to make more accurate predictions across different subgroups of people.

Personality has been shown to be reliably predictive of many different behaviours (Paunonen, 2003). Despite this, the interest in using personality to predict consumer behaviour is relatively new, and in part has grown due to the inability of demography, culture, and social class to adequately explain or predict purchasing behaviour (Solomon et al., 2012). Using personality traits, over motivations and attitudes, is advantageous as just five trait dimensions can be used to predict behaviour over different contexts and periods in time. Personality theory hypothesises that people who are more similar on a trait value will behave more similarly than those who are different, in any given situation (Tkalcic et al., 2009).

Introducing The Big Five Personality Traits

One of the ways personality can be theorised is in terms of traits. Traits are defined as “a continuum of various personality characteristics” (Burger and Caldwell, 2000). Trait theory is the idea that individuals have dispositional, stable, and enduring mechanisms that influence their behaviour, thoughts, and decisions (Asendorpf, 1992). The measure of these traits in combination with one another constitutes a personality. Trait theory is empirically useful as it allows personality to be measured quantitatively using scales.

A five factor structure of personality traits has arisen from factor analyses of both descriptive personality adjectives and sentence-based questionnaires. This structure is called the Big Five (or the Five-Factor Model) (Costa and MacCrae, 1992; Digman, 1990; Goldberg, 1990; John et al., 1999). The same analyses in languages other than English produce slight variations in these five constructs, and it is possible an additional domain may improve the generalisability of the taxonomy across different cultures (Ashton et al., 2004; Saucier and Goldberg, 2001). Nonetheless, the five factor structure has shown to be robust and useful at predicting many different behaviours across multiple domains including health, education, psychiatric disorders, life expectancy, salary, and consumption. The Big Five is a useful tool as it provides a common language for researchers to study the effects of individual differences on different behaviours.

The five traits are commonly referred to as Agreeableness (compassion and cooperative), Extraversion (sociable and assertive), Conscientiousness (hard-working and organized), Openness (creative and intellectual), and Neuroticism (anxious and susceptible to low or unstable moods) (Costa and McCrae, 1985; McCrae and John, 1992). These five over-arching domains of personality sit at the top of the personality framework hierarchy, where trait facets make up the lower levels, and specific behaviours are at the bottom (McCrae and John, 1992; Paunonen, 1998). The five dimensions are bi-polar so that opposite ends of the scale reflect different characteristics. For example, the opposite of extraversion (sociable and excitable) is introversion (quiet, and preferring to be alone).

Openness

Individuals who are high in Openness (O) are imaginative, artistic, cultured, and happy to go against the status quo. Those low in O, are conventional, conservative, dogmatic in their attitudes and values and have narrow interests (Costa Jr et al., 1991; Lakhal et al., 2012). The lexical model of personality (Goldberg, 1992) labels O as ‘Intellect’, aligning with the finding that O is correlated with intelligence (DeYoung et al., 2005; Ackerman and Heggestad, 1997). Whilst the questionnaire approach recommends an ‘Openness’ label, analogous to curiosity. Of the five factors, O typically shows the lowest correlations between lexical and questionnaire measures (Goldberg, 1992). More recently O has been conceptualised as comprising of two related yet slightly variant facets - openness to new experiences, and intellect (McCrae, 1993; DeYoung et al., 2009).

Conscientiousness

Those high in Conscientiousness (C), are characteristically self-disciplined, reliable, punctual, ambitious, organised, competent, and diligent individuals who display planned rather than spontaneous behaviour and are more likely to be perfectionists (Zurawicki, 2010). Whilst those low in C are disorganised, unmotivated, and lackadaisical (Costa and McCrae, 1992).

Extraversion

Extraversion (E) (also spelled as extroversion) is the extent to which people primarily obtain gratification from external sources (Costa and McCrae, 1992). Individuals high in E typically enjoy human interactions, and partaking in social activities more than those low in E. They are confident, enthusiastic, energised, talkative, assertive, and seek stimulation. Whilst individuals low in E are introverted, quiet, shy, and happy to be in their own company (Costa and McCrae, 1992).

Agreeableness

Agreeableness (A) is an expression of the need for harmonious relations (Zurawicki, 2010). Those high in A tend to be trusting, sympathetic, cooperative, good natured, forgiving and better at emotional self-regulation (Ho et al., 2004). Low A individuals are argumentative, cynical, callous, and antagonistic (Costa and McCrae, 1992).

Neuroticism

Neuroticism (N) refers to individual differences in negative emotional response to threat, frustration, or loss (Costa and McCrae, 1992; Goldberg, 1992). Individuals high in N find it harder to control or stabilise moods, making them more susceptible to negative emotions such as irritability, anger, sadness, anxiety, worry, hostility, self-consciousness, and a lack of self-esteem (Costa and McCrae, 1992; Goldberg, 1992). Such negative emotional responses are both more frequent, and out of proportion to the circumstances (McCrae and Costa, 2003). Conversely, those low in N, are emotionally more stable, calmer, have higher self-esteem, and generally happier.

The Big Five and Consumer Behaviour

Personality has been shown to predict different consumption patterns. Interest in personality as a predictor has grown in part due to the inability of demography, culture, and social class to adequately explain or predict purchasing behaviour (Solomon et al., 2012). Using traits, as opposed to motivations and attitudes, to predict behaviour is advantageous as personality remains relatively stable over time and contexts (Costa and McCrae, 1980). In addition, data collection is more efficient when using traits, as just five key constructs are able to predict a multitude of general behaviours.

Extraversion particularly has been linked to different consumption behaviours. One study looking at consumers' feelings and attitudes towards purchasing cars found that extraversion and neuroticism correlated strongly with customer satisfaction and post-purchase behaviour (Mooradian and Olver, 1997). Extraversion has also been shown to predict status purchasing, the buying of flashy or expensive items to indicate wealth (Wang and Cheng, 2013), even when income is accounted for (Landis and Gladstone, 2017). Other consumption behaviours where differences have been explained by personality include, impulse buying, compulsive buying, and hedonic buying - each of which, and their relationship with different traits, is described below in more detail.

The Big Five and Impulse Buying

People high in C are able to control impulses and delay gratification better than people low in C (Joshani et al., 2012). This ability may explain the negative relationship between impulsive behaviour and conscientiousness found by Gustavsson et al. (2003). Duijsens and Diekstra (1996) reported a negative relationship between agreeableness and impulsive behaviour, but George (2002) did not find sufficient evidence to support this claim. Khan et al. (2016) similarly found that extraversion had no effect on impulse buying in fashion retail.

The Big Five and Compulsive Buying

Compulsive shopping (also known as oniomania) is an abnormal form of consumer behaviour typified by "chronic buying episodes...in which the consumer feels unable to stop or significantly moderate the behaviour(s)" (Bearden and Netemeyer, 1999). Compulsive buying differs from impulsive buying in that it has been classified by the American Psychological Association as an (under-recognized and hard-to-treat) disorder. Compulsive buying disorder is characterised by excessive or poorly controlled preoccupations, urges, or behaviours regarding shopping and spending (Black, 2001). Compulsive consumption has been linked to traits such as dependence, materialism, perfectionism, denial, depression, approval seeking, low self-esteem (d'Astous, 1990), excitement seeking, and lack of impulse control (DeSarbo and Edwards, 1996) - many of which are typical of a high N individual. Indeed, Johnson and Attmann (2009), found a significant relationship between compulsive buying and neuroticism.

Similarly, significant positive correlations have been found between compulsive buying and depression; anxiety and nervousness (Valence et al., 1988); and other lifetime mood disorders (Black et al., 1998). In addition to N, C (Gohary and Hanzae, 2014; Mowen and Spears, 1999) and O (Gohary and Hanzae, 2014) have also been found to predict compulsive buying, as well as impulsive buying and utilitarian shopping values.

The Big Five and Hedonic Buying

Hedonic buying is the purchasing of goods for pleasure and luxury. Matzler et al. (2006) studied the relationship between the hedonic value of products people bought and their personality traits. Matzler et al. found that people high in extraversion and openness are more likely to buy highly hedonic products. The study also showed that openness (directly) and extraversion (indirectly via hedonic value) influenced brand affect (Chaudhuri and Holbrook, 2001). Various other studies have shown a positive relationship between hedonic buying motives and openness, extraversion, and neuroticism - where consumers who scored high on these traits sought fantasies (Tsao and Chang, 2010), excitement (Holbrook and Hirschman, 1982), fun (Strahilevitz and Myers, 1998), and enjoyed online shopping (Kim and Shim, 2002).

3.1.3 Impulsivity

Impulsivity, traditionally defined as “a tendency to act rashly or without adequate forethought” (Adams et al., 2012), features in all major theories of personality (Whiteside and Lynam, 2003). Trait impulsivity is one's predisposition to act rashly across different situations. This predisposition can be moderated by self-control or other strategies. Thus, someone can have an impulsive predisposition but not act on their impulses. There is consensus in the literature that impulsivity is multifaceted (Whiteside and Lynam, 2003; Dickman, 1990; Cyders et al., 2007) and that behavioural measures (for an overview see (Cyders and Coskunpinar, 2012) may not relate strongly to trait impulsivity ($r=0.097$ (Cyders and Coskunpinar, 2012)).

Generalised measures of impulsive tendencies (and a lack of them) are the BIS/BAS scales, measuring the strength of two opposing internal systems: the Behavioural Inhibition System (BIS) and the Behavioural Approach System (BAS) (Carver and White, 1994). The BIS is responsive to stimuli associated with punishment, extreme novelty (Corr et al., 1995), high intensity, and innate fear stimuli. Typical behaviours influenced by the BIS are stopping any actions, scanning the environment, and passive avoidance. BIS is typically accompanied by feelings of fear and anxiety. The reactivity of the BIS system differs between individuals, influencing their proneness to ‘trait anxiety’ (Carver and White, 1994).

The second system is the Behavioural Approach System (BAS). This system is responsive to incentives and reward (Corr et al., 1995). The feelings accompanying BAS activation include excitement and anticipation. A heightened sensitivity of the BAS influences an individual's impulsivity. The BAS construct is broken down into three dimensions: Reward Responsiveness (BAS-RR), how sensitive a person is to the pleasure derived from, or in anticipation of, rewards; Drive (BAS-D), which measures a person's level of motivation towards a goal; and Fun Seeking (BAS-FS) which captures the need to experience novel and exciting rewards and experiences.

Impulsivity and consumer behaviour

Impulsivity is related to consumer behaviour through shopping impulsively, or impulse buying. Shopping impulsively is the expression of an impulsive nature within the consumer context, and is outlined more extensively in Section 3.1.6. Individuals who are impulsive can express their inherent rash nature in different domains, but in general if someone has higher trait impulsivity, they have a higher disposition to shop impulsively (Whiteside and Lynam, 2003). However, this disposition can be ameliorated by other traits of influences such as income, self-control, future focus, conscientiousness, frugality, and credit rating. Thus, trait impulsivity and shopping impulsivity are measured separately here.

3.1.4 Self-Control

Self-control (SC) can be defined as “the ability to suppress inappropriate emotions, desires, and actions in favour of alternative, appropriate ones” (Casey, 2015), for example delaying gratification (Mischel et al., 1988). Stable individual differences in SC have been found using self-report measures. Prominently, Tangney et al. (2004a) Self-Control Scale (SCS) measures SC as a trait. The SCS items assess impulse control, emotional regulation, self-discipline, performance management, and ability to break bad habits. The SCS has been shown to predict various positive outcomes, such as academic grades (Wolfe and Johnson, 1995), implying SC is an adaptive trait. Tangney et al. (2004a) showed that SC positively predicted the strength of interpersonal relationships, family cohesiveness, self-acceptance, and self-esteem. However, it was also predictive of negative psychological symptoms such as somatization (the manifestation of physical symptoms to psychological stress), obsessive-compulsive patterns, paranoia, anxiety, hostile anger, and depression (Tangney et al., 2004a).

Self-Control and Consumer Behaviour

Self-control (SC) can influence consumer behaviour because of its influence on impulsive buying (i.e. Hoch and Loewenstein, 1991). Higher levels of SC suppresses the impulsive and hedonistically-driven urge to buy. It is also likely that SC will predict receptiveness to different kinds of marketing strategies. For example, Baumeister (2002) theorised that someone low in SC will be more vulnerable to being “seduced by the moment”, and therefore advertisements which emphasize immediate gratification are likely to be more appealing and successful. In contrast, someone high in SC may be more persuaded to buy something if the long-term benefits and value are emphasised. Furthermore, (Romal and Kaplan, 1995) found evidence linking SC to frugality measures: people high in SC managed their finances better, saved more, and spent less.

3.1.5 Time Preferences

Time preference refers to individual differences in the value placed on future outcomes relative to immediate ones. Time preferences are trait-like in that they’re acquired early on,

stable, and affect behaviour (Fuchs, 1980). Those who have a greater present-focus are more likely to ‘live in the moment’, be impulsive, emotionally reactive, and have less patience. People who are more future oriented are better at planning ahead, contemplating long-term consequences, less ‘hot-headed’, and have greater self-control.

Time preferences are commonly measured using Multiple List Price (MPL) questions (Holt and Laury, 2002) which ask the respondent a series of questions such as “would you rather have £20 now or £50 in two weeks?”, to quantify at what point the present bias (wanting money now) switches over to choosing a greater future payout. Time preferences have been shown to influence a number of factors including mental health, refusal to report weight or height, snacking, smoking, binge drinking, seat belt use, completing homework, exercise, health insurance, installing energy-efficient lighting, thermostat temperature, participation in welfare programs, retirement savings, and credit card balance (Bradford and McAlvanah, 2014; Ariely and Wertenbroch, 2002; DellaVigna and Malmendier, 2006; Fang and Silverman, 2009; Ruhm, 2012).

Time Preferences and Consumer behaviour

Economic theories predict that present-focused individuals are less likely to spend money, time, or resources in the present, to yield benefits in the future (Bradford and McAlvanah, 2014). Thus, time preferences are inherently linked to saving and spending patterns with present focused individual more likely to engage in impulse buying, whilst those who are more future focused more likely to be frugal and save money. Chapman and Elstein (1995) suggested that an individual’s’ time preference may vary depending on the domain. However, Bradford and McAlvanah (2014) found that time preferences elicited using real monetary pay-outs were significantly correlated with many outcomes related to health, energy use, and finances. Moreover, they found that time preferences on monetary outcomes were a greater predictor of actual health outcomes, than time preferences elicited from hypothetical health questions.

Time preferences are particularly relevant to the health domain as present-focused individuals are more likely to underestimate the effects of short term behaviours such as eating too much, exercising too little, and smoking, on long term consequences (Bradford and McAlvanah, 2014). Fuchs (1991; 1996) found that differences in time preference can help explain variations in these health-related choices. Various studies have found connections between time preference and BMI (Chabris et al., 2008; Sutter et al., 2013), exercise (Chabris et al., 2008; Bradford, 2010), smoking (Bradford, 2010; Sutter et al., 2013), alcohol consumption (Sutter et al., 2013; Bishai, 2001), preventive health care utilisation (Bradford, 2010), snacking and binge eating (Bradford, 2010) healthy behaviours among hypertensive patients (Axon et al., 2009), and overall self-assessed health (Van der Pol, 2011). Blaylock et al. (1999) asserts the reason behind poor dietary habits in America (despite extensive information on health and nutrition) is due to Americans heavily discounting the future.

Consumption-specific Variables

3.1.6 Shopping Impulsivity

An impulsive purchase is defined by (Piron, 1991) as “an unplanned purchase, result, and instantaneous decision of a consumer’s response to a stimulus.” It is also typical that after the purchase the consumer “experiences emotional and/or cognitive reactions”. Impulse buying makes up between 40% (Hausman, 2000) and 80% (Abrahams, 1997) of all purchases in some goods categories. 38% of adults self-report as an impulsive buyer (DDB Needham annual Lifestyle Survey, 1974-93). And 87% of American adults admit to making impulse purchases (NEFE, 2010). Impulsive buying generates over \$4 billion in annual sales in the US (Kacen and Lee, 2002). Thus stimulating impulsive buying can create a large competitive advantage for shops and businesses. The greater anonymity and convenience of online shopping makes online impulse buying more prevalent. According to LaRose and Eastin (2002) one quarter to four fifths of online consumers are engaging in impulse buying.

The predisposition to buy impulsively can be considered a facet of general impulsiveness (Beatty and Ferrell, 1998; Gerbing et al., 1987; Puri, 1996; Weun et al., 1998) as well as its own unique trait (Verplanken and Herabadi, 2001; Weun et al., 1998). Recently the consensus has swayed in favour of a distinct purchase impulsivity trait (Whiteside and Lynam, 2003). Consumers who are high in purchase impulsivity characteristically have more ‘open’ shopping lists, and are more receptive to unexpected buying ideas (Rook and Fisher, 1995). Impulsive buyers are unreflective in their buying cognitions, are prompted by physical proximity to a desired product, succumb more often to a product’s affect and attraction, and seeking in immediate gratification (Hoch and Loewenstein, 1991; Thompson et al., 1990).

Impulsive buyers have the impulse to buy, stronger and more frequently, than non-impulsive buyers due to their predisposition. Moreover, because they reflect less, impulsive buyers are inherently more likely to respond affirmatively to their buying impulses. In extreme cases, impulsive buying is almost entirely stimulus driven, translating directly into a physical response, with no behavioural control (Levy, 1976; Solnick et al., 1980). A further defining characteristic of impulsive buying first alluded to by (Piron, 1991) is the accompaniment of post-purchase feelings of guilt, or an ignorance/denial of future consequences (i.e. the consideration of money problems).

Not helped by conceptualisations of impulsive buying as ‘a lack of control’, the general attitude to impulsive buying is that of a socially unacceptable phenomenon (Virvilaite et al., 2015). Consumer’s attitudes and normative evaluations of impulse purchasing are important, as they moderate the relationship between trait-like shopping impulsiveness and actual buying behaviour (Rook and Fisher, 1995). Therefore, retailers try to make impulsivity seem desirable, to encourage impulsive purchasing. Public conceptions of impulsive purchasing as ‘irresponsible’ and ‘immature’ (Tinne, 2010) has provoked alterations to self-report questionnaires. For example Rook and Fisher (1995) replaced ‘impulsive buying’ with ‘spontaneous buying’ in attempts to alleviate negative connotations and reduce social desirability effects.

Contrary to the negative stereotype, Nesbitt (1959) argued that impulse buying is the more intelligent way to shop. By being open and flexible people can take advantage of in store offers, and maximising purchasing power.

Various different motivations behind impulsive purchasing have been researched. People may impulse buy to satisfy hedonistic demands (Rook and Fisher, 1995), to reward themselves if they are in a good mood (Rook and Fisher, 1995), to attempt to alleviate a negative emotional state such as depression (Korgaonkar and Bellenger, 1980), to bridge the gap between perceived self and ideal self (Sirgy, 1982, 1985), or to cheer themselves up (Mick and DeMoss, 1990). Motivations may or may not reach consciousness. In spite of varying motivational pathways, items which bring a pleasurable state/emotion to the customer are the main target for impulsive buyers. Emotions are such an intrinsic part of impulse buying, Virvilaite et al. (2015) described the moment of purchase as the customer being “controlled by emotional forces”.

Koski (2004) showed that the following factors increase impulsive buying: ease of accessibility, anonymity, variety of goods, direct marketing, use of credit cards, immediate satisfaction from product, how hard it is to return to the store, a good feeling of the purchasing environment, and the inability to compare prices and products. Parboteeah (2005) added that greater impulsive buying occurs when more time is spent in the store, when there are greater opportunities to examine the products, when there is a positive symbolic meaning of the goods, when products are discounted or on sale, and in supermarkets. Various findings have related demographics to impulse buying. Women have been found to be greater impulsive buyers than men (Kollat and Willett, 1967; Coley and Burgess, 2003; Parboteeah, 2005; Dittmar et al., 1995; Rook and Hoch, 1985); younger people feel less risk when spending, making them more inclined to buy impulsively (Parboteeah, 2005); and between the ages of 18 and 39 impulse buying increases slightly, and thereafter declines (Wood, 1998; Bellenger et al., 1978).

In addition to demographic differences, particular personalities have been shown to indicate a vulnerability to impulse buying. Trait impulse buyers are on average lower in conscientiousness, have low autonomy, a low personal need for structure, a reduced need to evaluate, are higher in extraversion, and highly action orientated (Verplanken and Herabadi, 2001). Additionally an impulsive buyer’s cognitive profile is more hedonic, more materialistic, and less utilitarian than a non-impulsive buyer (Verplanken et al., 2005; Podoshen et al., 2014).

Measures of Trait Impulsive Purchasing

An accurate measure of impulsive buying behaviour is important for understanding the impact of trait impulsivity. Having just the measure of the predisposition is not satisfactory, as the two are not synonymous. A person may have a high disposition to buy impulsively, yet not act on their impulses. Trait impulsive purchasing has most commonly been measured using by Rook and Fisher (1995) 9 item self-report scale, modelled on Eysenck’s trait

impulsivity. Other methods however exist and include Weun et al. (1998) 5 item scale, and Verplanken and Herabadi (2001) 20 items measure. Actual impulsive buying behaviour has been measured in various ways: using retrospective self-report questions such as “How often have you bought things spontaneously in the last month?” (Rook and Fisher, 1995), virtual projection tasks i.e. “What would Mary do?” (Rook and Fisher, 1995), or using field studies (Rook and Fisher, 1995; Beatty and Ferrell, 1998).

Self-report questions involve participants recalling instances of impulse purchasing from long-term memory, and thus accounts are subject to unconscious biases and heuristics (Tversky and Kahneman, 1974; Taft, 1954). Projection tasks, although they reduce social desirability, assume that what the participant chooses for someone else, they would also choose for themselves. Additionally, asking what someone should/would do in the virtual situation, is different to measuring what someone will actually do in that situation. Field studies, although measuring real behaviour, are confounded by sampling bias. For example, there’s a certain type of person who will stop and offer to fill out surveys at the end of their shopping trip. Field studies can also require recalling plans and intentions to buy, prior to the shopping experience. These can be internally confused and replaced in memory by things actually brought, or thoughts occurring during the shopping experience. Beatty and Ferrell (1998) field study used pre and post shopping interviews to reduce the reliance on memory. However contemplating planned items immediately prior to shopping, and the knowledge that someone will be inspecting your purchases directly after, is likely to influence behaviour. Furthermore, previous research, such as Rook and Fisher (1995) field study, failed to account for social visibility. Luo (2005) showed that shopping with others increases the likelihood of an impulse purchase. Thus although a reliable and valid measure of trait impulsivity has been established, no such equivalent yet exists for actual impulsive buying behaviour. Hence, in part what motivated this research, particularly the linking of a personality questionnaire to transaction data, was to investigate whether loyalty card transaction data could be used as a measure of impulsive buying.

3.1.7 Frugality

There has been an increasing recognition among consumer researchers about the importance of understanding, not just consumption in consumer behaviour, but non-consumption as well (Gould et al., 1997). Frugality, defined as “the degree to which one is both restrained in acquiring and resourceful in using products” (Bove et al., 2009), is an important influence on consumption behaviour. Whether frugality is a trait (Guber, 2003), or better described as a set of values (Todd and Lawson, 2003), a lifestyle type (Lastovicka et al., 1999), or a pattern of behaviour (e.g., Egol et al., 2010) is disputed. The debate mirrors that categorising materialism (Belk, 1984; Richins and Dawson, 1992) which is often regarded as the converse of frugality (Todd and Lawson, 2003).

Bove et al. (2009) found that older people were more likely to be frugal, but found no relationship between frugality and religiosity, or social class. Lastovicka et al. (1999) showed that extraversion (negatively), agreeableness (positively) and conscientiousness (positively) correlated with frugality measures, whilst Bearden et al. (2006) found that frugal consumers are more future focused in their time preference. Todd and Lawson (2003) found that people lower in frugality scored higher on the values: pleasure, enjoying life, exciting life, varied life, detachment, social recognition and true friendship – all values which have been linked to extraversion. People who were frugal reported their important values as: honesty, politeness, obedience, and responsibility – values typically associated with agreeableness.

Frugality and Consumer Behaviour

Frugality is inherently linked to consumption. Frugal people are characterised as being less materialistic (Goldsmith et al., 2014; Pepper et al., 2009), buying less (Rose et al., 2010), less subject to social influences (Lastovicka et al., 1999; Goldsmith et al., 2014), more ecocentric (Dunlap and Van Liere, 1978), more price consciousness (Shoham and Brenčič, 2004) less involved with brands (Goldsmith et al., 2014), and are more likely to repair and re-use items (Albinsson et al., 2010). Conserving behaviours can stem from both satisfaction in resourcefulness as well as ecological motivations (Dunlap and Van Liere, 1978). Rick et al. (2007) highlights that frugal people are distinct from “tightwads”, as they enjoy saving money as opposed to disliking spending it. Pepper et al. (2009) found that frugal behaviour was related to income constraints, however Lastovicka et al. (1999) showed that frugal behaviours occur independently of income.

Saucier and Goldberg (1998) found frugality appeared within a cluster of person descriptors, that could not be explained by the Big Five, suggesting it is an independent trait-like construct. However, there is also evidence to suggest that frugality is context dependent. Research shows that when the economy is strong, frugality fades (Witkowski, 2003). Todd and Lawson (2003) describe frugality as a “fashion idea that is cyclical: it comes in and out of style”. However Goldsmith et al. (2014) argue that although external forces can influence frugal behaviour, frugality is “a syndrome of individual differences that are independent of economic conditions”.

3.1.8 Brand Personality and Brand Commitment

Through the introduction of the now widely used concept of brand personality (Aaker, 1997) personality and brands are inherently linked. Brand personality refers to particular brands seeming to take on human traits. Aaker (1997) created a brand personality scale, measuring the following five dimensions of brand personality: sincerity, competence, excitement, ruggedness, and sophistication. Brand personalities give consumers something with which they can relate to, and identify themselves with. Studies have shown that consumers are inclined to buy brands that reflect their own personalities (Lee, 2009). Brand personality

has also been shown to affect the type and strength of commitment consumers show to the brand (Louis and Lombart, 2010). Both consumer's brand commitment and brand loyalty have been associated with individual differences in risk aversion (Matzler et al., 2006). For example, the more risk averse a consumer is, the more loyal and committed they are to a brand, both behaviourally and attitudinally (Day, 1976). Studies have shown that women tend to be more risk averse than men (Byrnes et al., 1999), and women also show higher levels of brand commitment (Tifferet and Herstein, 2012).

Marketing scholars Sheth and Parvatiyar (1995) theorise brand loyalty is a tactical move for consumers to reduce perceived risk. Matzler et al. (2008) showed that the relationship between risk aversion, brand loyalty and brand commitment is mediated by brand trust and brand affect. When brand affect and brand trust were included in Matzler et al.'s model, the previously strong relationship between (domain specific) risk aversion and attitudinal brand loyalty was no longer significant. Therefore, brand affect and brand trust, which are also related to brand personality (Sung and Kim, 2010), are likely to be important influences on brand commitment.

3.1.9 Summary

This literature review has identified many different ways in which demographic and psychological variables may relate to consumer buying behaviour. The construction of the final questionnaire was informed by this literature. The survey needed to be long enough to provide detailed quality information; yet short enough to avoid a large drop-out rate (which would select a sample of people who were more likely to persevere). A further limitation on what variables could be measured in the questionnaire, was the availability of reliable open-source measurement scales. The final questionnaire can be viewed in full in the Supplementary Materials (Section 10.1.1). The next section describes the data from the survey data, as well as the respondents' transaction data.

3.2 The Data

The previous section outlined the literature which fed into a crucial part of data collection for this thesis - the questionnaire. The next section introduces the self-report data that was attained from that survey, along with the loyalty card transaction data to which it was linked.

3.2.1 Loyalty Card Transaction Data

Descriptives

The Boots Advantage Card scheme currently has around 14.4 million active members (having used their card within the last 6 months) (Boots, 2019). The loyalty card is one of the most generous reward programmes, offering four points for every pound spent in store (Baldwin,

Table 3.1: Descriptive statistics for 10% sample of Boots’ 10 year transaction dataset

	Mean	SD	Min (≥ 0)	Max
Spend per person	£836.31	1,412	£0.00	£200,299.29
Items bought per person	218.22	405.75	1	28,688
No. of transactions (baskets) per person	61.9	100.39	1	4,097
Spend per basket	£13.51	21.32	£0.00	£11,454.50
No. of items per basket	3.53	3.67	1	1,933
Spend per item	£9.63	19.33	£0.00	£2,960.85

2018). The loyalty card data received for this research consists of a 10% sample of the whole of Boots loyalty card data spanning 10 years from 17/04/2005 to 16/04/2015. The data is structured and stored in multiple tables in a database accessible using SQL database language. The data includes what customers bought which items, when they bought them, how much they cost (and other item information), as well as where they were bought from (with other store information). Data was only acquired when a customer bought something and their loyalty card was scanned (to accumulate points) in the same transaction. This loyalty card number, and the individual’s information attached to it, allows for all transactions to be linked to the person, so their behaviour can be studied over time. The dataset contains 2,702,449 distinct customers, 234,938,722 distinct transactions (baskets) in which there were 718,105,269 transaction lines and 744,991,152 items bought. The data’s descriptive statistics can be found in Table 3.1.

3.2.2 Survey Data

Members of Boots’ customer panel completed an online survey, which included self-reported measures of personality and demographic questions. They provided permission - through informed consent - for researchers to access four years of their loyalty card transaction data. The transaction data was pseudo-anonymized to link self-reports with individual’s loyalty cards number. No incentives were offered for taking part in the study. This study was reviewed by the School of Computer Science, University of Nottingham ethics committee, and full ethical approval was attained prior to surveys being distributed.

Survey questionnaires were sent to 80,000 members of the retailer’s internal consumer panel. 12,968 were returned fully completed and matched to participants’ purchasing data. The questionnaire comprised of questions on age, gender, current occupation, annual household income, highest education qualification, how many people they live with, how many children do they have, marital status, time preferences (Reimers et al., 2009), self control (Tangney et al., 2004b), a frugality measure (Lastovicka et al., 1999), impulsivity (Carver and White, 1994), the Big Five (Gosling et al., 2003), a shopping impulsivity scale (Rook

and Fisher, 1995), questions on shopping motivations, and (self-reported) geographical region and shopping frequency.

The Ten Item Personality Inventory (TIPI) (Gosling et al., 2003) was chosen to measure the Big Five as a short yet reliable measure of personality was required. The TIPI also has norm information attached to it, which could be used to compare to the averages of the survey respondent’s personality traits. Impulsivity was measured using the Behavioural Activation Scales (BAS) (Carver and White, 1994). The BAS incorporates a measure of Reward Responsiveness (the level of excitement associated with attaining a reward) (BAS-RR); a measure of Drive (BAS-D) (the persistence towards a goal); and Fun Seeking (BAS-FS) (both a desire for new rewards and the tendency to approach potentially rewarding experiences without careful consideration). Due to space restrictions on the survey, the BIS measurement was not included, as the interest was primarily in approach not the avoidance of stimuli in a consumer setting.

3.2.3 Survey Data Descriptives

Table 3.2: The descriptive statistics for the continuous variables for all completed surveys ($N=13,189$)

	Mean	SD	Min	Max	No. of items	Scale
Age (years)	50.82	14.93	17	93	-	-
O	4.85	1.19	1	7	2	1-7
C	5.84	1.08	1	7	2	1-7
E	4.04	1.54	1	7	2	1-7
A	5.49	1.12	1	7	2	1-7
ES	4.7	1.46	1	7	2	1-7
BAS	2.74	0.46	1	4	11	1-4
BAS-RR	3.2	0.46	1	4	5	1-4
BAS-FS	2.76	0.52	1	4	4	1-4
BAS-D	2.42	0.64	1	4	4	1-4
Self Control	3.33	0.64	1.08	5	13	1-5
Shopping Impulsivity	2.7	0.86	1.11	5	9	1-5
Frugality	4.61	0.87	1	6	8	1-6
Happiness	7.16	2.15	0	10	1	0-10
Satisfaction	7.18	1.93	0	10	1	0-10

Note: O= Openness, C= Conscientiousness, E= Extraversion, A= Agreeableness, ES= Emotional Stability, BAS= Behavioural Activation System, BAS-RR= Reward Responsiveness, BAS-FS= Fun Seeking, BAS-D= Drive

Of the 13,189 completed questionnaires (12,968 of which were successfully matched to transaction records) 11,746 were female (89%), 1,152 were male (8.7%), and 291 chose not to answer the question (2.2%). The respondents were from Great Britain and Northern Ireland. The survey asked the respondents to choose the region where they live: 502 were from East Anglia, 526 from the South West, 552 identify from Northern Ireland, 562 from London, 618 from Wales, 623 from the West Midlands, 665 from the North East, 765 from Scotland, 922 from the East Midlands, 1,050 from the North West, 1,165 from the South East, and the largest number ($N=1,513$) were from Yorkshire.

The distribution of the income groups are shown in Figure 3.1. These were: ‘Less than £25,000’ ($N= 3,002$), ‘£25,000 to £34,999’ ($N=2,175$), ‘£35,000 to £49,999’ ($N= 2,264$), ‘£50,000 to £74,999’ ($N= 1,876$), ‘£75,000 to £99,999’ ($N= 743$), ‘£100,000 to £149,999’ ($N= 383$), ‘£150,000 or more’ ($N= 169$), and ‘Prefer not to say’ ($N= 2,577$).

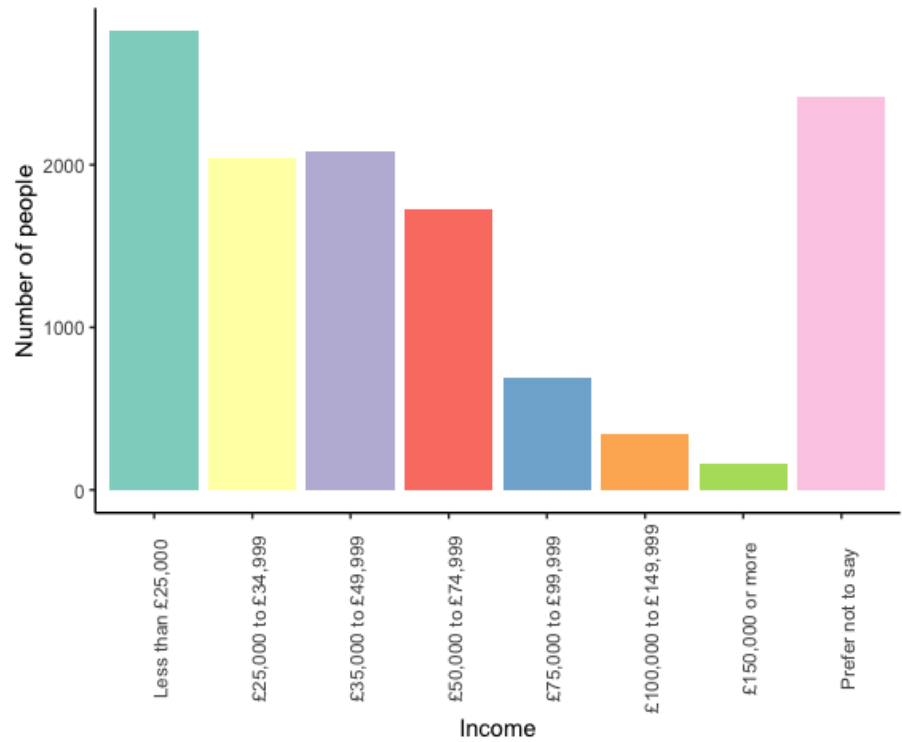


Figure 3.1: The survey respondents income.

The questionnaire asked about respondents marital status (see Figure 3.2). The majority ($N= 8,953$) were married, 2,188 were single, 1,180 were divorced, 585 were widowed, 226 were separated, and 57 answered ‘other’. When people were asked about the people in their household, the majority ($N= 5,704$) responded that they lived with a partner, 2,438 lived with a partner and children, 2,351 lived on their own, 2,049 lived with other adults that were family, 356 lived with no partner and children, and 291 people lived with adults that were not family members. Of those that had children, most had one child ($N= 1,401$), 1,118 had two children, 204 had three, 43 had four, and six respondents had five children.

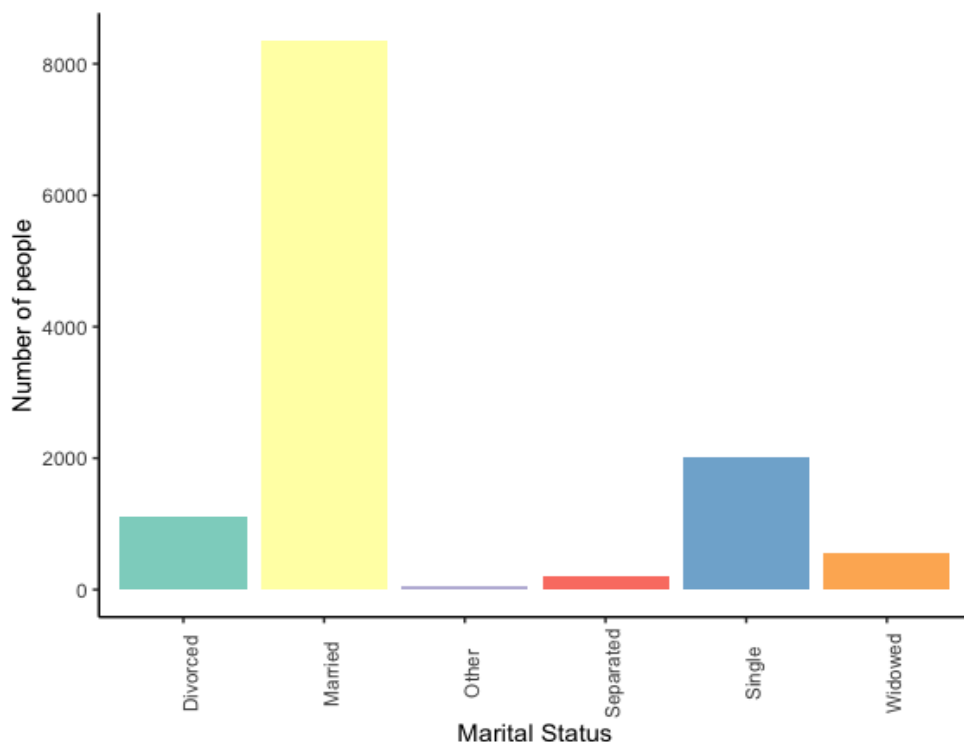


Figure 3.2: A graph to show the survey respondents marital status”

Respondents reported their occupation type as one of the following types (see Figure 3.3): “Intermediate managerial/ professional/ administrative” ($N=3,466$), “Supervisory/ clerical/ junior managerial/ professional/ administrative”; ($N=2,807$), “Higher managerial/ professional/ administrative” ($N=2,367$), “Retired/ Semi-retired” ($N=1,596$), “Skilled manual worker” ($N=1,200$), “Semi/ unskilled manual work” ($N=564$), “Other” ($N=369$), “Unemployed/ long-term sickness” ($N=260$), “self-employed/ own business” ($N=152$), “Homemaker” ($N=144$), “Full-time carer of other household member” ($N=102$), “Student” ($N=92$), or “Casual worker” ($N=70$).

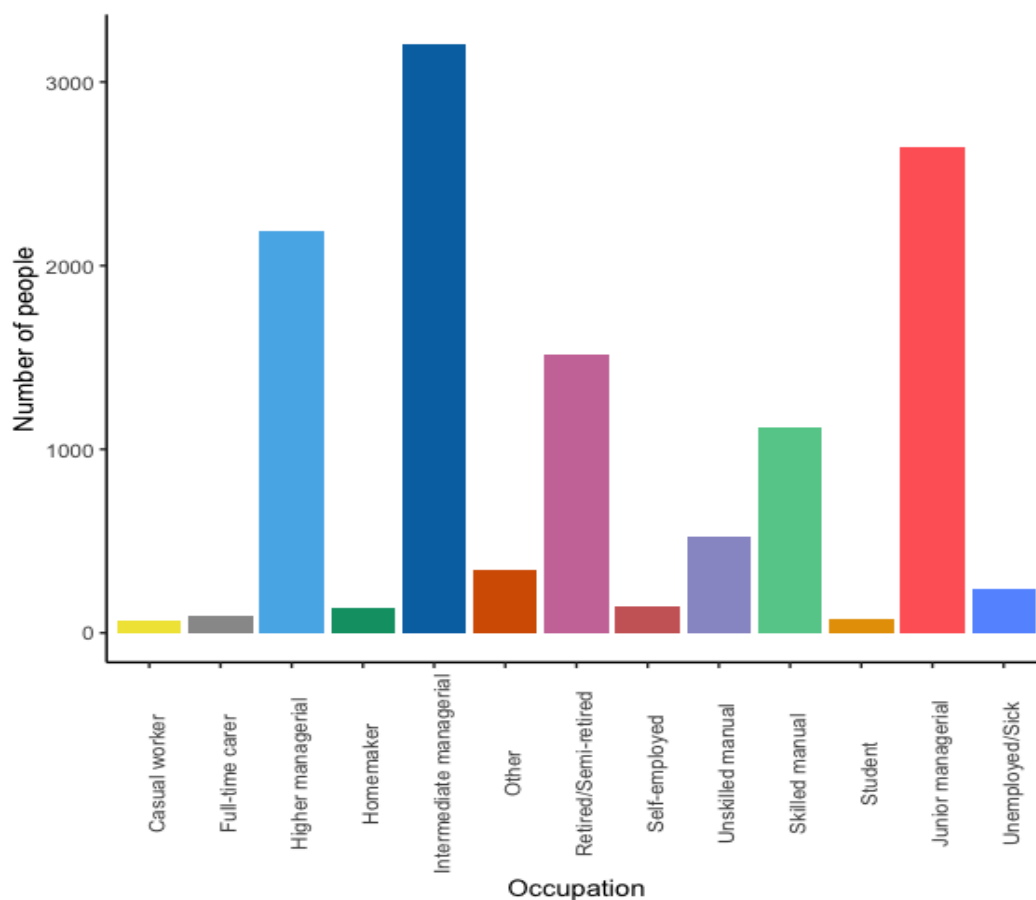


Figure 3.3: A graph to show the survey respondents occupation type. ‘Junior managerial’ refers to “Supervisory/ clerical/ junior managerial/ professional/ administrative”; and ‘unskilled manual’ refers to “Semi/ unskilled manual work.”

The respondents were asked what their highest level educational qualification was (see Figure 3.4). The majority of respondents highest qualification was a Bachelor’s degree ($N=2,966$), 2,587 had GCSE’s or equivalent, 2,421 had A-levels or equivalent, 1,555 had ‘other’ postgraduate qualifications (i.e. a postgraduate diploma), 1042 had masters, 772 said their highest qualification wasn’t listed as an option, 703 had a BTEC or equivalent, 613 people reported completing some secondary school, 327 had a PhD, law or medical degree, 162 had AS-level as their highest, and 41 people reported having some other advanced degree beyond a masters.

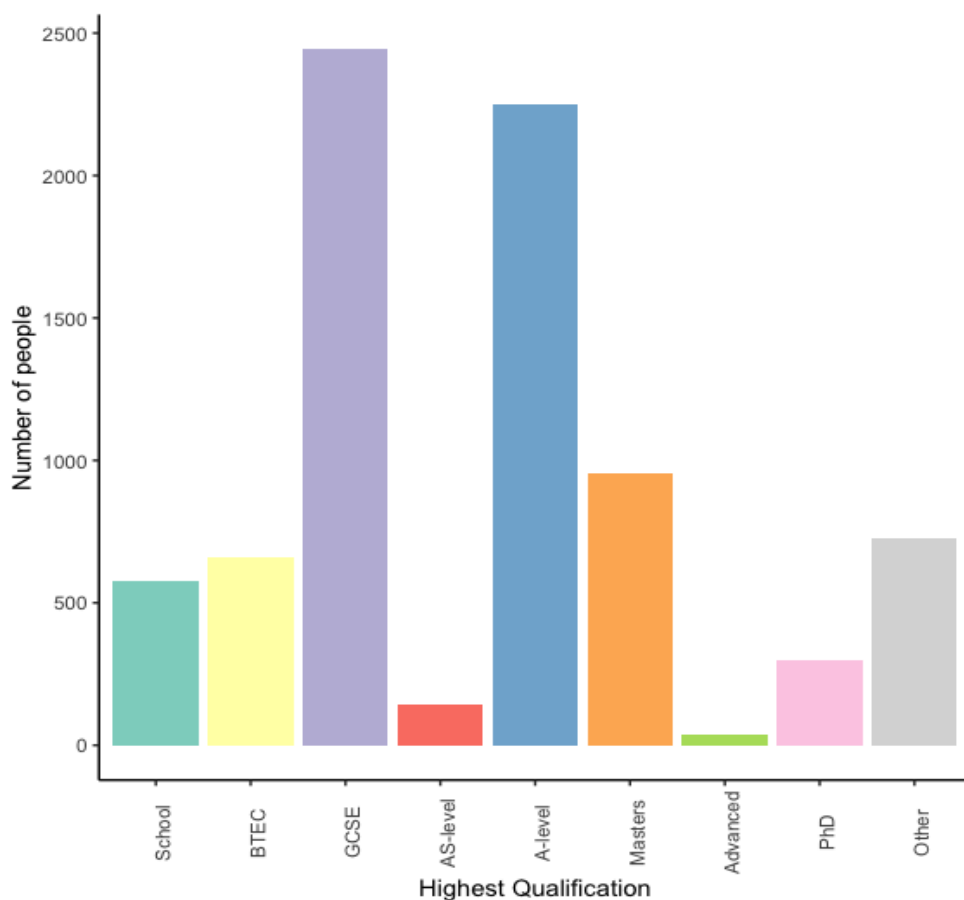


Figure 3.4: A graph to show the survey respondents highest level of qualification

The survey also asked respondents to self report how often they shopped at Boots, and how often they used their Boots Advantage Card (see Figure 3.5). The majority of people reported shopping once a fortnight ($N=3,946$), 3,448 reportedly shopped at Boots once a month, 2,969 once a week, 1,072 once every 2-3 months, 800 twice a week, 639 more than twice a week, 142 once every 4-6 months, 79 every 6 months or less, 55 every day, and 39 people reported not knowing how often. Figure 3.6 shows that most people used their advantage card every time they brought something at Boots ($N=11,270$), 1,532 people reported to using their card most of the time, 301 said sometimes, 74 people responded hardly ever, and 12 people said they never use their advantage card.

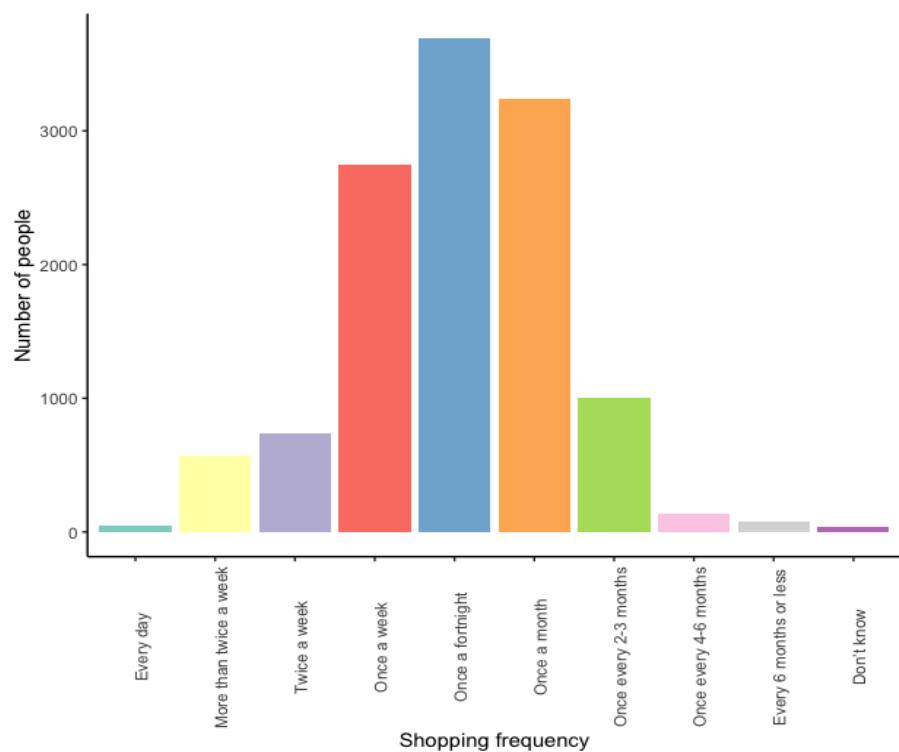


Figure 3.5: A graph to show how often people reported shopping at Boots

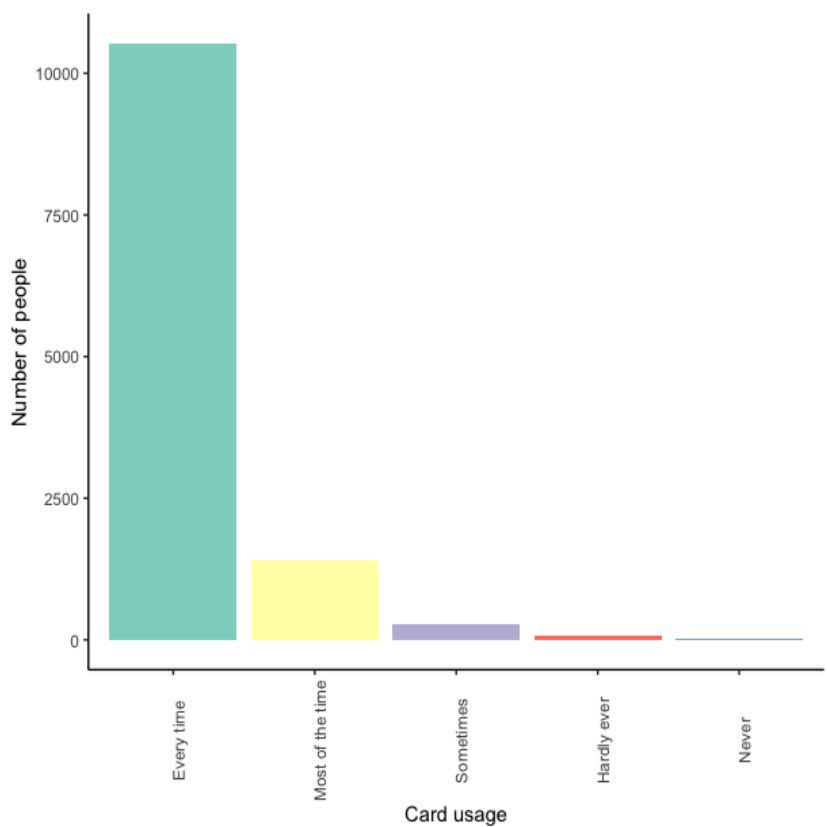


Figure 3.6: A graph to show how often people reported using their loyalty card at Boots

The time preference questions were ‘If offered the following financial alternatives which would you rather have?’: £25 now or £70 in 3 months, £40 now or £70 in 3 months, £55 now or £70 in 3 months. The distribution of answers can be viewed in Figure 3.7.

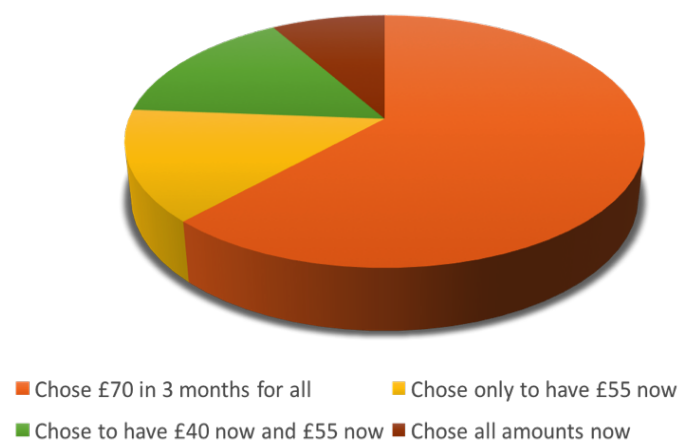


Figure 3.7: Graph to show the distribution of answers across the time preference questions.

The descriptive statistics for the continuous psychological variables can be found in table 3.2. The Big Five personality scores were measured using the Ten Item Personality Inventory (TIPI) Gosling et al. (2003). Figure 3.8 shows the comparison of the mean big five traits from the Boots survey data compared with the TIPI norms for each age group (Gosling et al., 2014). The level of extraversion is very similar, and not varying much with age. Agreeableness is slightly higher in the Boots sample compared to the norms, and following the same trend of increasing in agreeableness with age. Conscientiousness is much higher in the Boots data, and increases with age to a lightly lesser degree than the scale norms. Emotional stability is a similar level in the Boots data compared with the norms, but varies slightly more with age - starting lower than the norms in the youngest age group, and finishing higher than the norms in the oldest age group. Finally, openness is considerably lower in the Boots data compared with the TIPI norms, and after the age of 30 doesn't drop off as it does in the norms. There was less data on males by age group, and so the mean personality scores across all age groups were compared to the TIPI means for males. The comparisons are shown in Figure 3.9. The males, like the females, had the same level of extraversion than the scale norms; were more agreeable; more conscientious; slightly more emotionally stable; and less open to new experiences. Overall, these descriptives show that the sample of questionnaire respondents with Boots loyalty card users were notably more agreeable, more conscientious, and less open to new experiences than in the general population.

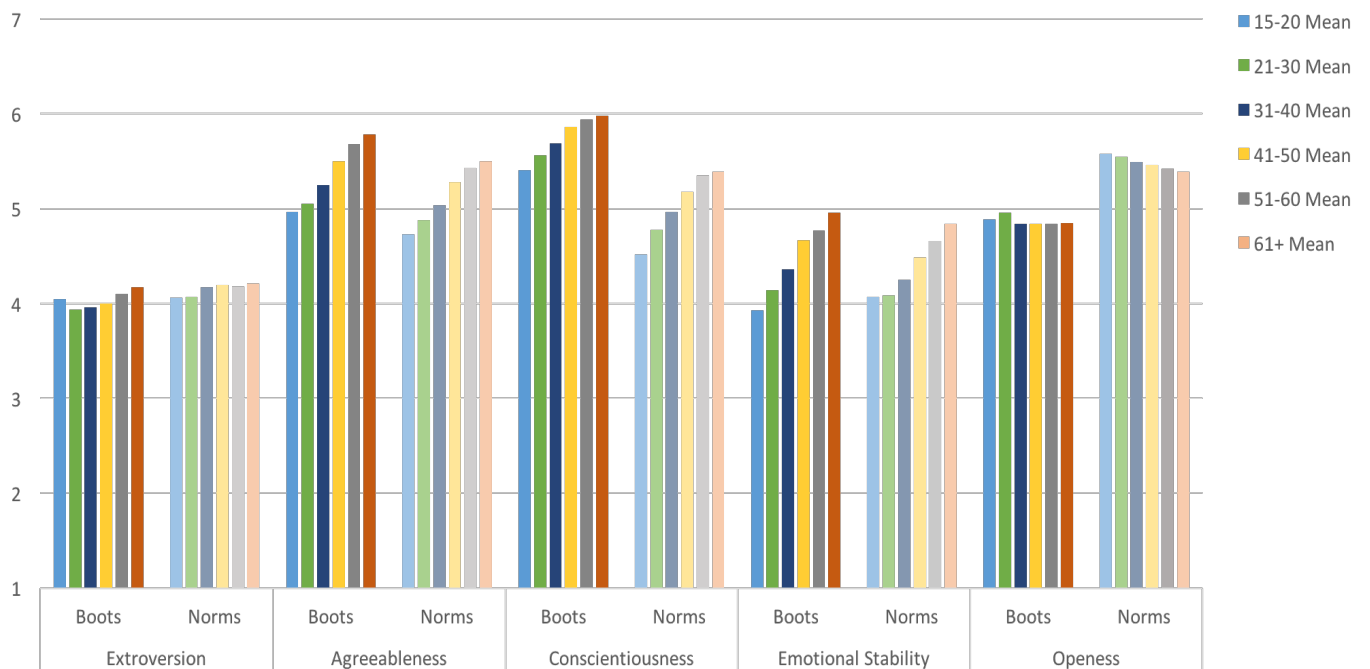


Figure 3.8: The mean trait scores for females in the Boots data compared with the female norms of the Ten Item Personality Inventory (TIPI) by age group (years).

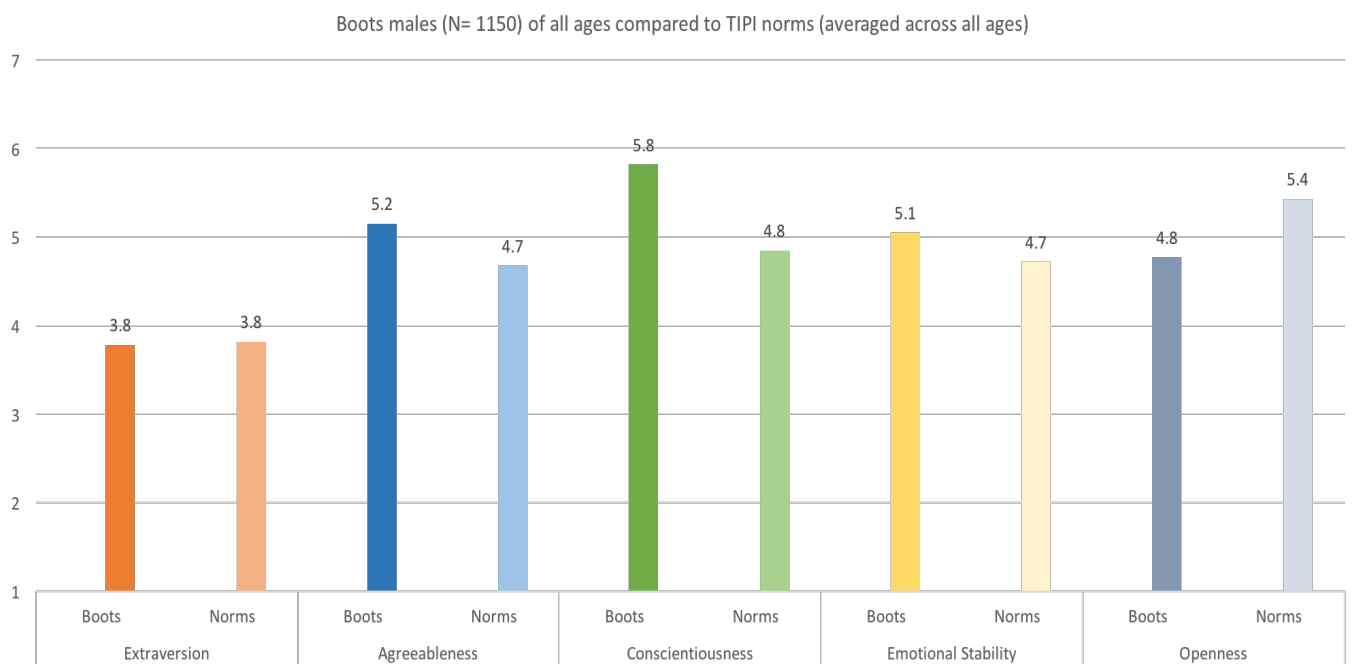


Figure 3.9: The mean trait scores for males in the Boots data compared with the male norms of the Ten Item Personality Inventory (TIPI). As there is less data on males for different age groups in the Boots data, the average across all age groups is used here for comparisons.

Table 3.4: Survey sample (4 year) transaction data descriptive statistics

	Mean	SD	Min (≥ 0)	Max
Spend per person	£1,680.04	1,531.46	£0.00	£47,897.45
Items bought per person	364.434	365.84	1	10,454
No. of transactions (baskets) per person	102.07	81.42	1	1,380
Spend per basket	£16.46	27.12	£0.00	£4,055.00
No. of items per basket	3.57	4.38	1	928
Spend per item	£10.27	20.03	£0.00	£2,695.92

Table 3.3: Statistics about the completion of the survey from all completed responses ($N= 13,189$)

	Mean	SD	Min	Max	No. of items	Scale
Relevance	7.59	1.67	0	10	1	0-10
Repetitiveness	3.46	2.42	0	10	1	0-10
Duration (Minutes)	23.52	238.04	2.18	24,241.68*	-	-

**Long duration's could have occurred if people left the survey open whilst completing other tasks and then coming back later to complete it.*

3.2.4 The Survey Sample’s Transaction Data

The survey sample’s transaction data is the transaction data of the customers who took the survey (extracted from 100% of the Boots data). This transaction data, however, only spans 4 of the 10 years from 25/05/2012 to 25/07/2016. Before removing customers with incomplete survey information the survey sample contains 15,530 distinct customers, 1,585,267 transactions (baskets) in which 5,658,515 items were bought.

Chapter 4

Personality and Consumer Behaviour Predicts Subjective Well-being

4.1 Introduction

Well-being is an increasingly important metric. It underpins the formation of new social policy (Duncan, 2005; Deeming, 2013) and its relevance continues to grow not only in psychological domains, but in economic and medical ones too (de Chavez et al., 2005). Improving understanding of the factors that influence well-being therefore remains of high interest; and, in particular, eliciting how an individual's personality, context and behaviours interact to increase or reduce it. Recent research has brought to light that increased spend and raw materialism alone are insufficient to generate happiness from money (Matz et al., 2016), and research is now focusing on *how* people are spending their money (i.e Zhang et al. (2014); Dunn et al. (2011)).

4.1.1 Money and Happiness

The relationship between money and happiness has been shown to be complex. Indeed, the extent to which money can improve an individual's well-being, beyond acquiring a basic comfortable lifestyle, remains unresolved. Stevenson and Wolfers (2013) analysed multiple datasets and found the relationship between well-being and income to be log-linear. In contrast, others have claimed that beyond fulfilling 'basic needs' money does not in fact improve well-being (Diener and Seligman, 2004; Clark et al., 2008; Di Tella and MacCulloch, 2008). Diener et al. (2002) suggested the mixed evidence on the interplay between money and happiness is because that relationship differs from person to person. They showed that an increase in individual's income can lead to highly variable happiness outcomes, suggesting that this variation was due to a myriad of different factors evoking happiness in different individuals. For example, Diener et al. (2002) found that people who value material goals over other goals are substantially less happy - unless they are rich. They provided evidence that there exist factors at play which moderate the relationship between money and happiness - in their case the value people place upon material goods.

Recent research on the money-happiness relationship has focused less on total income, and more on what people specifically spend their money on. Dunn et al. (2011) show that if money is spent in the ‘right’ way, it can improve people’s well-being. They suggest 8 ways in which people can get more happiness from their money, including: buying more experiences rather than material goods (Carter and Gilovich, 2010; Van Boven and Gilovich, 2003); using money to benefit others (Dunn et al., 2014); and buying many small pleasures as opposed to a few large ones (Dunn et al., 2011). This research suggests that money does have the potential to improve well-being, but that to do so people may need to spend it on goods that are apposite fits to them as individuals.

Personality is key to what makes people different, and the individual differences literature has provided a rich illustration of the utility of using personality traits to predict well-being (Diener et al., 2003). The Big 5 model of personality (Goldberg, 1990), for example – which posits that people’s personalities are well described by five traits: openness to experience (artistic vs. conservative), conscientiousness (self-controlled vs. easy going), extraversion (outgoing vs. reserved), agreeableness (compassionate vs. antagonistic), and neuroticism (emotionally stable vs. unstable) – has underpinned numerous studies which have investigated the effect of personality on well-being. Neuroticism (N) is consistently found to be a strong negative predictor of well-being (Vittersø, 2001; Costa and McCrae, 1980; Veenhoven, 1984; Hayes and Joseph, 2003; Gutiérrez et al., 2005; Quevedo and Abella, 2011). Extraversion (E) has been found to be a positive predictor (Costa and McCrae, 1980; Hayes and Joseph, 2003; Gutiérrez et al., 2005; Quevedo and Abella, 2011), in particular through facets such as social support (Lu, 1997). Hayes and Joseph (2003) found that Conscientiousness (C) is positively related to well-being, in particular satisfaction with life. Agreeableness (A) is not considered in the literature to be strongly related to well-being, however it has been found to predict positive relations with others, often conceived as a facet of well-being (Schmutte and Ryff, 1997). Some studies have found Openness to New Experience (O) to be positively linked to well-being (Hayes and Joseph, 2003; Quevedo and Abella, 2011), whilst others have found O to be positively related to depression (Wolfestein and Trull, 1997). This contradiction can be explained via the suggestion that people high in O have more feelings of both positive and negative affect (Costa Jr et al., 1991).

Recent research has focused on the thesis that individual differences can moderate the effect of purchases on happiness. For example, experiential purchases improve happiness, but more so for experiential people (Zhang et al., 2014). Gifting can improve happiness for those who sympathise highly with others’ concerns, but not for those who are less agreeable (Hill and Howell, 2014). Proto and Rustichini (2015) found that traits moderate the effect of income on life satisfaction. They found the effect is strongest for N, with people high in N enjoying their income more if they are poorer, and less if richer. When this income-N interaction was introduced, the significant main effect of income disappeared. In particular, Matz et al. (2016) found that people report experiencing more happiness from their money if they purchase things that were designated as corresponding to their personality. Extroverts

experienced more positive affect from spending money on drinks in a bar than introverts; and introverts experienced more positive affect from purchasing a book than extroverts. This literature highlights the importance of using an individual difference approach when studying the ways in which money makes people happy.

Individuals have also been shown to actively choose different environments to suit their personality, known as ‘psychological-fit’ (Sirgy, 1985), which subsequently can positively affect their well-being. Jokela et al. (2015) found this effect for people living in London boroughs, where people’s personalities matched the personalities of those in their neighbouring boroughs more closely than those in boroughs further away. They also found that the degree of psychological-fit people had to their neighbours positively predicted well-being. In a different study, Matz et al. (2016) used bank transaction data linked to a personality questionnaire measuring well-being and the Big 5, to examine the effect of psychological-fit to people’s purchases, and the relationship this had with well-being. First, products were rated on the Big 5 personality traits, treating them as if they were people. These ratings were then consequently used to calculate individuals’ psychological-fit to their shopping. They found that people purchase more from the product categories which better fit their personality, supporting the evidence provided by Sirgy (1985). Additionally, Matz et al. (2017) found that when people have better psychological-fit with their shopping basket that this positively predicted well-being better than total spending or income. Therefore, they concluded that money can indeed engender happiness - if it is spent on the things that match your personality.

4.2 Matz et al. (2016), a Replication: Buying things that match your personality makes you happier

4.2.1 Introduction

Large transaction data-sets for research, especially those with personality information attached to them, are not common in psychological research. This makes replications of big data psychology studies of this nature rare. The Boots data provided a unique opportunity to investigate whether the findings from Matz et al. (2016), that buying in line with your personality makes you happier, are replicable. If a finding is robust it should be able to be found in new data, and even using variant methods. This research replicates and extends the findings from Matz et al. (2016) that the degree to which peoples’ personalities match their purchases (psychological-fit) positively predicts well-being.

Matz et al. (2016) used bank transaction data linked to a personality questionnaire measuring well-being and the Big 5, to examine the effect of psychological-fit to people’s purchases, and the relationship this had with well-being. Matz et al. (2016) got Amazon’s Mechanical Turks to rate products on the Big 5 personality traits, treating them as if they were people. The instructions given were - “We would like you to think of each (spending)

category as if it were a person... This may sound unusual, but think of the set of human characteristics associated with each spending category” (Matz et al., 2016). It may “sound unusual” because it is unusual to personify inanimate objects or spending categories. Brands have long been studied in relation to the human personality traits they encompass. Brand personalities, a set of ‘human characteristics associated with a brand’ (Aaker, 1997) have allowed the estimation of similarity between an individual’s personality and what they buy. Yet Avis et al. (2014) found that if people are asked to rate rocks on human personality traits, humans will be able to do it and statistical differences emerge. Just because people personify inanimate objects, doesn’t render it meaningful. This study uses the product category ratings in Boots loyalty card data to see whether the same effect can be recovered.

As well as performing a replication, the current study extends the work by Matz et al. in two ways. Firstly, it looks not just at overall basket-participant (b-p) match (encompassing all of the Big Five traits together) but it also studies the match on each personality trait separately. It was predicted that the extent to which basket-person match predicted well-being would vary for different traits. This premise is supported by recent research that found that the psychological fit on trait openness for people living in London boroughs predicted well-being, but not for any of the other traits (Jokela et al., 2015) Therefore, the study extends Matz et al.’s research to investigate which traits are important to match on for happiness in a consumer context. This will provide a greater understanding of what traits underlie the relationship between basket-person match and well-being.

The second way this study extends the work of Matz et al. is to pilot a new method of calculating psychological-fit which negates the need for human ratings. As the task of rating products as having personalities is quite arbitrary, the current study investigates whether these ratings can be derived from the data instead. It is proposed that products reflect the personality of the people who buy them. Thus, how product categories differ in their psychological nature can be quantified through calculating the average personality of the people who buy them. The b-p match is calculated using 2 methods: first, using the human crowd-sourced ratings from Matz et al. (2016); and second, using a data-derived b-p metric. This study uses both methods to predict well-being, evaluating their comparative performance.

4.2.2 Method

Participants

Participants were from the sample of customers who’s completed survey responses were matched to their loyalty card data, as described in Chapter 3, Section 3.2. When the participants volunteered their data as part of the online survey, it was made clear that they were giving permission for their pseudo-anonymized data to be linked to their personality questionnaire prior to full anonymization. Of the 12,968 respondents, 12,408 people had bought from at least one of the 10 product categories used in this study (see Materials section). 20%

($N = 2,435$) of these people had answered ‘prefer not to say’ to the income question, and so these people were left out of the analysis. Therefore the final dataset consisted of 9,933 people, and 2,474,011 individual transactions. Of these 9,010 were female (91%) and 923 were male (9%). The mean age was 50 years ($SD = 14.93$). No incentives were offered for taking part in the study. This study was carried out in accordance with the recommendations of the School of Computer Science, University of Nottingham ethics committee with informed consent from all participants. All participants gave informed consent in accordance with the Declaration of Helsinki.

Materials

The electronic personality survey (for the full version see Section 10.1.1 in the Supplementary Materials) included the Ten-Item Personality Inventory (TIPI), measuring the Big 5 (Gosling et al., 2003) on a 7-point likert scale. Well-being was measured using two questions: “Overall, how satisfied are you with your life nowadays?” and “Overall, how happy did you feel yesterday?” both on a scale of 0-10. In line with Longo et al. (2016), the happiness and satisfaction measures were highly correlated ($r = 0.90$), and so were combined using the mean into one well-being variable. The 8 income categories were transformed to a continuous income variable by coding the categories from 1 to 7, with 1 being ‘less than £25,000’ and 7 being ‘£150,000 or more’. Gender was dummy coded with males = 0 and females = 1. See Table 4.1 for descriptives and correlations between variables.

The products represented within the raw transactional data were grouped into 10 categories that matched the Matz et al. (2016) category descriptions. Those that did not fit with Matz et al.’s category descriptions were not used in this analysis. The 10 product categories used for this study were: travel (number of people who had bought at least one item from the category ($N=7,672$), snacks (‘bakers and confectioners’ in Matz et al. (2016)) ($N=6,588$), books ($N=1,369$), dental care ($N=8,896$), hair and beauty ($N=9,887$), health and fitness ($N=7,087$), home furnishing ($N=1,804$), jewellery ($N=1,976$), photography ($N=3,904$) and toys and hobbies (not including children’s toys) ($N=1,086$).

Calculating Psychological-fit

To calculate psychological-fit, a personality score of both the person and the products they are purchasing is required. Both this study and the Matz et al. (2016) study attained the person’s personality scores from a questionnaire. Where the two methods of calculating psychological-fit differ is how the products’ personalities are measured.

Method 1: Matz et al. (2016) attained the product personalities by recruiting 100 Amazon Mechanical Turks to rate their categories according to how the category name was associated with each of the Big Five personality dimensions. The raters were asked “if this ‘spending category’ were a person, how would it best be described?”. These ratings are therefore capturing people’s perceptions of a product’s personality in an anthropomorphic fashion. Intra-class correlation coefficients ranged from .82 to .98, indicating good inter-rater

agreement. These 10 product category ratings were extracted directly from the Matz et al. (2016) paper for use in this study. They were originally rated on a 7 point scale from -3 to 3 using the Big Five Inventory-10 (BFI-10) personality questionnaire (Rammstedt and John, 2007), and were subsequently converted to a 1-7 point scale to align with the ratings from the TIPI personality scales used in the personality questionnaire measuring customer’s personality traits. Although the customer’s personality and the product categories were measured on different scales, the scales have high convergent correlations (Gosling et al., 2003).

Method 2: As well as replicating Matz et al. (2016)’s method of getting humans to rate product categories, this section also investigates an alternative way to associate personality traits with people’s shopping. This method derives the product personality scores from the data. These are calculated as the mean trait scores of every person who had bought from that category. If a person had purchased an item from the product category more than once, their duplicated personality scores were not removed. This meant that the product personalities were weighted towards people who bought more frequently from that product category. Therefore this calculation of psychological-fit is measuring whether people conform to the same purchasing habits as others with a similar personality to themselves, a contrasting approach to the more arbitrary attribution of personalities to products that Matz et al. (2016) employ. Figure 4.1 provides a comparison of the mean centred z-standardised category trait scores from Matz et al. (2016)’s study, and the data derived method. This illustrates the distinct difference between data-driven ratings to those assigned to products in the Matz et al. (2016) study. It also evidences that the two methods are measuring conceptually different aspects of consumer behaviour.

Once both sets of product scores (using Method 1 and Method 2) had been computed, Z-scores for both the customer’s and the product’s personality traits were calculated. To calculate an overall shopping or ‘basket’ personality score, the mean trait score was taken from the product categories in which the person had bought a product in. This was done for each of the Big Five. As in Matz et al. (2016), on the basis of research showing that many small purchases can result in greater happiness than a few large ones (Dunn et al., 2011), equal weights were assigned to each spending category regardless of how much was spent in them. The Euclidean distance (Hernández et al., 2011) was then calculated for each trait between the z-score of the subject (i) and their basket (b). Psychological-fit or the basket-participant (b-p) match was calculated as the square root of the sum of the Euclidean distances for each trait, and subtracting it from the mean (see 4.1 where O refers to trait Openness.) This calculation was undergone for each of the Big Five personality traits: first using Matz’s product personality ratings, and second using the proposed novel method of deriving product ratings from the data.

$$basket-participant match_{i,b} = mean - \sqrt{(z(O_i) - z(O_b))^2} \quad (4.1)$$

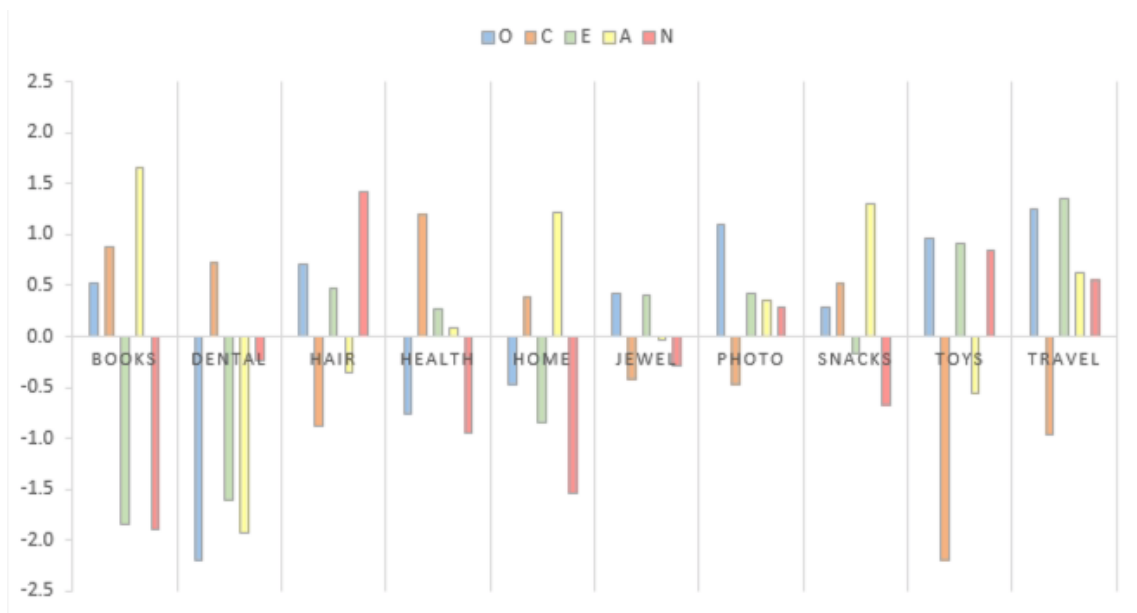
Multiple Linear Regression Analysis

For both of the methods for calculating an individual's b-p match enumerated above, 3 different multiple linear regression models were formulated in order to assess the cumulative effect of demographics, personality traits, and b-p match on well-being. Model 1 consisted of gender, age, log transformed ($\log_e$) total spend (the total amount spent in the 10 categories) and income. As in Matz et al. (2016) study, all the continuous variables were standardized and mean-centred before being input into the model. Model 2 included the addition of the personality traits of the customer, and the product category personalities (as controls). Two further sub-models were then created in order to analyse the additional variance explained by b-p match. Model 3a included b-p match calculated with all the five traits together as in Matz et al. (2016). Model 3b was created to investigate which traits are the most important to match on, and inputted the b-p match of each of the Big 5 separately.

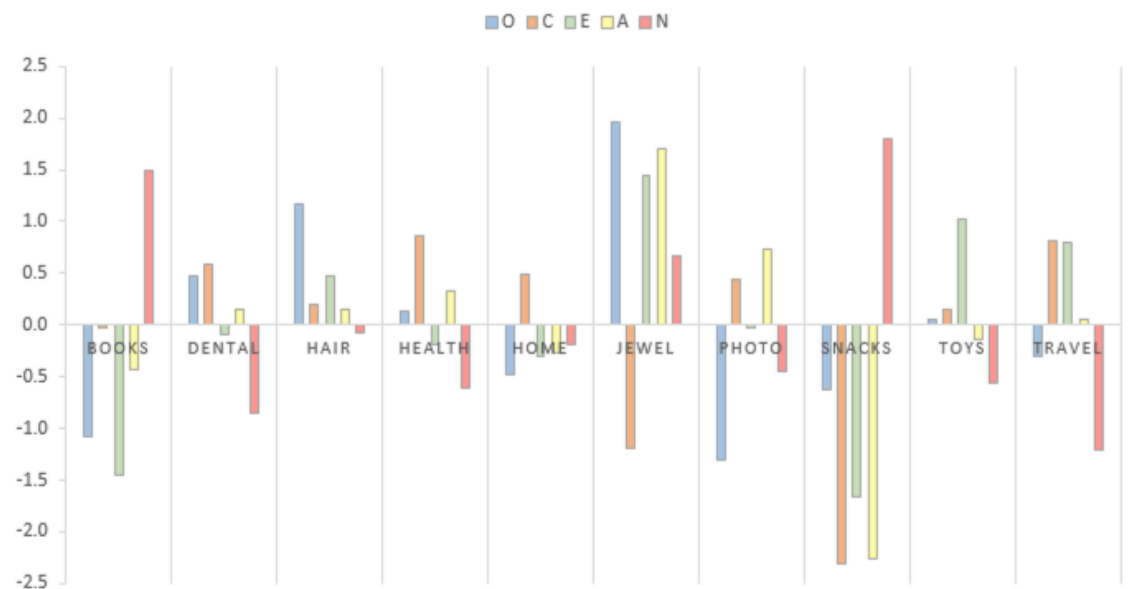
4.2.3 Results

Table 4.2 first shows the results of the Multiple Linear Regression (MLR) models predicting well-being using Matz et al.'s method of calculating psychological fit. In Model 1, all the demographic variables (age, gender and income) and total spend were statistically significant positive predictors of well-being, and the model explained 4.9% of the variance in well-being. In Model 2, all the person Big 5 variables significantly predicted well-being. N had the largest effect size, negatively predicting well-being. O also negatively predicted well-being. The other traits (C, E and A) positively predicted well-being. Together the traits added an additional 15.1% variance explained. As expected, the personalities of the products were not predictive of well-being.

In Model 3a the combined b-p match variable was a statistically significant ($p < 0.05$) positive predictor of well-being, replicating the findings of (Matz et al., 2016). When the traits were added separately in Model 3b, b-p match on O was a negative predictor of well-being; b-p match on C was not significant; and b-p match on E, A, and N were significant positive predictors of well-being. N accounted for the greatest proportion of the b-p match effect. Both Models 3a and 3b accounted for 20% of the variance (adding an additional 0.002 for 3a and 0.003 for 3b variance explained). Both statistically improved the prediction of well-being from Model 2.



(a) Method 1: Human ratings



(b) Method 2: Data derived

Figure 4.1: Graphs to compare the standardised Z-scores (on the y axis) of the personality traits for the 10 product categories used in this pilot study. The categories are books, dental care, hair and beauty, health and fitness, home furnishing, jewellery, photography, snacks, toys and hobbies, and travel.

Table 4.3 shows the results using MLRs to predict well-being using the study’s novel data derived method of calculating psychological-fit. Model 2 varied very little to the Matz et al. replication, with the only distinct difference being that the product O score is a negative predictor of well-being, where previously it was not. N was again the strongest predictor, and along with O negatively predicted well-being. The other personality traits significantly positively predicted well-being. Together the traits added an additional 15.1% variance explained to the model. In Model 3a the b-p match variable was a statistically significant positive predictor of well-being. In Model 3b, b-p match on O was a significant negative predictor of well-being, and b-p match on N was a significant positive predictor. The b-p match on the other traits were not significant. Both Model 3a and 3b explained 20% of the variance in total (adding an additional 0.0016 variance explained for 3a and 0.0013

		Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12
1	Age	50.13	-14.93												
2	Gender: Female	0.91	-0.29	-0.15											
3	Total Spend (£)	979.83	-949.08	-0.04	0.15										
4	Income	2.71	-1.51	-0.2	-0.03	0.22									
5	O	4.84	-1.18	-0.03	0.03	0.02	0.03								
6	C	5.82	-1.09	0.13	0	0.04	0.08	0.06							
7	E	4.06	-1.55	0.03	0.06	0.07	0.1	0.35	0.07						
8	A	5.48	-1.12	0.22	0.08	0.04	-0.07	0.14	0.2	0.06					
9	N	3.32	-1.47	-0.2	0.08	-0.01	-0.08	-0.2	-0.27	-0.22	-0.29				
10	Happiness	7.1	-2.15	0.16	0.01	0.04	0.1	0.09	0.18	0.18	0.15	-0.39			
11	Satisfaction	7.13	-1.93	0.13	0.01	0.05	0.14	0.08	0.2	0.19	0.15	-0.38	0.79		
12	Wellbeing	7.11	-1.93	0.15	0.01	0.05	0.13	0.09	0.2	0.2	0.16	-0.41	0.95	0.94	

Table 4.1: Descriptive statistics and the correlations between variables

for 3b). Both models 3a and 3b were a significant improvement from Model 2.

Model 3a and 3b using Matz et al.’s method and the data derived method of calculating psychological fit, both explained the same amount of variance (20%). However, due to the different pattern of variables, and the differences in the product’s personality scores, it is likely that the variability explained using this study’s data-derived method of calculating psychological fit, is different variability to that explained by (Matz et al., 2016).

4.2.4 Discussion

In line with Sirgy (1982) self-congruity model and the findings from Matz et al. (2016), the MLR models showed that the degree of psychological-fit between a person and their spending predicts well-being (when demographics and personality variables are controlled for). When the b-p match variable was added in Model 3a, it was a better predictor of well-being than total spend. Therefore, this research adds to the body of work which suggests that when it comes to money and happiness, it is not how much you spend, it is what you spend your money on that matters. Specifically, purchasing what people with the same personality as you purchase is an important positive predictor of well-being. Unlike the finding from Matz et al. (2016), the b-p match variables were not more important at predicting well-being than income in the regressions.

The addition of personality traits in Model 2 allowed the MLR to explain substantially more variance compared with using just demographics and total spend alone. This confirms the importance of the use of personality traits when predicting well-being (Gutiérrez et al., 2005). In line with previous research, N negatively predicted well-being (Costa and McCrae, 1980; Veenhoven, 1984; Hayes and Joseph, 2003; Gutiérrez et al., 2005; Quevedo and Abella, 2011), whilst E (Costa and McCrae, 1980; Veenhoven, 1984; Hayes and Joseph, 2003; Gutiérrez et al., 2005; Quevedo and Abella, 2011) and C did positively (Hayes and Joseph,

	Model 1 (F(4,50264)=644***, $R^2 = 0.049$ ***)			Model 2 (F(14,50254)=896***, $\Delta R^2 = 0.151$ ***)			Model 3a (F(15,50253)=844***, $\Delta R^2 = 0.002$ ***)			Model 3b (F(19,50249)=673***, $\Delta R^2 = 0.003$ ***)		
Predictor	β	95% CI	t	β	95% CI	t	β	95% CI	t	β	95% CI	t
Total Spend (log)	0.02	[0.01,0.03]	4.87***	0.02	[0.01,0.02]	3.60***	0.01	[0.00,0.02]	3.10***	0.01	[0.01,0.02]	3.38***
Gender: Female	0.03	[0.02,0.04]	7.25***	0.04	[0.03,0.05]	8.81***	0.14	[0.10,0.17]	8.50***	0.04	[0.03,0.04]	8.44***
Age	0.19	[0.18,0.19]	41.39***	0.09	[0.08,0.10]	20.15***	0.09	[0.08,0.10]	20.68***	0.09	[0.08,0.10]	20.35***
Income	0.16	[0.15,0.17]	34.32***	0.1	[0.09,0.10]	22.49***	0.1	[0.09,0.10]	22.48***	0.1	[0.09,0.10]	22.64***
Person:												
Openness				-0.03	[-0.04,-0.02]	-7.08***	-0.03	[-0.04,-0.02]	-6.35***	-0.03	[-0.04,-0.02]	-6.35***
Conscientiousness				0.08	[0.08,0.09]	19.92***	0.09	[0.06,0.08]	16.64***	0.09	[0.06,0.08]	16.72***
Extraversion				0.11	[0.10,0.12]	25.37***	0.11	[0.10,0.12]	25.34***	0.11	[0.10,0.12]	25.30***
Agreeableness				0.03	[0.02,0.04]	6.31***	0.03	[0.02,0.04]	6.58***	0.03	[0.02,0.04]	6.53***
Neuroticism				-0.33	[-0.34,-0.32]	-73.52***	-0.33	[-0.34,-0.32]	-73.77***	-0.33	[-0.34,-0.32]	-73.81***
Product:												
Openness				0	[-0.04,0.04]	-0.05	0	[-0.04,0.03]	-0.11	0	[-0.07,0.07]	-0.12
Conscientiousness				-0.02	[-0.04,0.02]	-0.82	-0.01	[-0.04,0.02]	-0.8	-0.01	[-0.07,0.03]	-0.83
Extraversion				0.01	[-0.01,0.02]	0.71	0.01	[-0.01,0.02]	0.73	0.01	[-0.02,0.04]	0.7
Agreeableness				0	[-0.03,0.03]	-0.1	0	[-0.03,0.03]	-0.07	0	[-0.06,0.06]	-0.05
Neuroticism				-0.01	[-0.03,0.02]	-0.57	-0.01	[-0.03,0.02]	-0.43	-0.01	[-0.06,0.04]	-0.43
Basket-participant match:												
b-p match on all traits tog.							0.05	[0.04,0.06]	9.46***			
b-p match Openness										-0.02	[-0.03,0.00]	-2.43*
b-p match Conscientiousness										0	[-0.02,0.02]	-0.01
b-p match Extraversion										0.05	[0.03,0.06]	6.54***
b-p match Agreeableness										0.02	[0.01,0.03]	2.81**
b-p match Neuroticism										0.08	[0.06,0.09]	10.92***

Table 4.2: Multiple linear regression models predicting well-being. Model 3a incorporates b-p match into one variable. Model 3b inputs the b-p match on each trait separately.

2003; Quevedo and Abella, 2011). N was the strongest predictor of well-being, more so than age, income, and E (supporting Vittersø (2001)). Contrary to Hayes and Joseph (2003), and Quevedo and Abella (2011), O significantly negatively predicted well-being.

Model 3b provided insight into which traits are important to match on, and which are redundant. With both methods of calculating psychological match, b-p match on N was the strongest positive predictor, despite trait N negatively predicting well-being. Therefore, if a neurotic person purchases things that are also deemed to be neurotic, this has a corrective effect on N’s negative relationship with well-being. Despite a person’s C score being an important predictor of well-being, b-p match on C was the only b-p match variable that was not a useful predictor of well-being using either method of calculating psychological-fit. B-p match on O negatively predicted well-being using both methods of calculating psychological fit. This shows that not all types of psychological-fit are positively related to well-being in a consumer context, contrary to what Matz et al. (2016) suggest. This result is set against the findings of (Jokela et al., 2015) where fit to people in London boroughs on O positively predicted well-being, highlighting that when predicting well-being psychological-fit to shopping acts differently to the psychological-fit to the surrounding population.

The results show that the novel method of calculating psychological-fit, using product

	Model 1 (F(4,50264)=644***, $R^2 = 0.049$ ***)			Model 2 (F(14,50254)=897***, $\Delta R^2 = 0.151$ ***)			Model 3a (F(15,50253)=846***, $\Delta R^2 = 0.0016$ ***)			Model 3b (F(19,50249)=666***, $\Delta R^2 = 0.0013$ ***)		
Predictor	β	95% CI	t	β	95% CI	t	β	95% CI		β	95% CI	t
Total Spend (log)	0.02	[0.01,0.03]	4.87***	0.01	[0.01,0.02]	3.51***	0.01	[0.01,0.02]	3.4***	0.01	[0.01,0.02]	2.51*
Gender: Female	0.03	[0.02,0.04]	7.25***	0.14	[0.03,0.05]	8.85***	0.14	[0.11,0.17]	8.68***	0.04	[0.03,0.04]	8.76***
Age	0.19	[0.18,0.19]	41.39***	0.09	[0.08,0.10]	20.16***	0.09	[0.08,0.10]	20.94***	0.09	[0.08,0.10]	20.53***
Income	0.16	[0.15,0.17]	34.32***	0.1	[0.09,0.10]	22.47***	0.1	[0.09,0.10]	22.55***	0.1	[0.09,0.10]	22.42***
Person:												
Openness				-0.03	[-0.04,-0.02]	-7.09***	-0.03	[-0.04,-0.02]	-6.90***	-0.03	[-0.04,-0.02]	-6.91***
Conscientiousness				0.08	[0.08,0.09]	20.03***	0.09	[0.08,0.09]	19.59***	0.09	[0.08,0.09]	19.58***
Extraversion				0.11	[0.10,0.12]	25.33***	0.11	[0.10,0.12]	25.51***	0.11	[0.10,0.12]	25.53***
Agreeableness				0.03	[0.02,0.04]	6.26***	0.03	[0.02,0.04]	6.08***	0.03	[0.02,0.04]	6.09***
Neuroticism				-0.33	[-0.34,-0.32]	-73.55***	-0.32	[-0.33,-0.32]	-72.51***	-0.32	[-0.33,-0.32]	-72.61***
Product:												
Openness				-0.02	[-0.03,0.00]	-2.70**	-0.02	[-0.03,0.00]	-2.54*	-0.02	[-0.03,0.00]	-2.54
Conscientiousness				0	[-0.03,0.02]	-0.04	0	[-0.03,0.02]	-0.03	0	[-0.02,0.03]	0.21
Extraversion				0.02	[-0.01,0.04]	1.49	0.02	[-0.01,0.04]	1.44	0.02	[0.00,0.04]	1.54
Agreeableness				-0.01	[-0.03,0.01]	-0.84	-0.01	[-0.03,0.01]	-0.75	-0.01	[-0.03,0.01]	-1.01
Neuroticism				-0.01	[-0.03,0.02]	-0.38	-0.01	[-0.03,0.02]	-0.48	0	[-0.03,0.02]	-0.21
Basket-participant match:												
b-p match on all traits tog.							0.04	[0.04,0.05]	8.79***			
b-p match Openness										-0.01	[-0.02,0.00]	-2.04*
b-p match Conscientiousness										-0.01	[-0.02,0.00]	-1.14
b-p match Extraversion										0.01	[0.00,0.02]	1.85
b-p match Agreeableness										0.01	[0.00,0.02]	1.62
b-p match Neuroticism										0.04	[0.03,0.05]	7.22***

Table 4.3: Multiple linear regression models predicting well-being. Model 3a incorporates b-p match into one variable. Model 3b inputs the b-p match on each trait separately.

personalities derived from the data, performed statistically the same as the crowd-sourcing method used by Matz et al. (2016). This is despite the product personality ratings exhibiting distinct differences (see Figure 4.1), suggesting they are indeed measuring different but equally explanatory aspects of psychological-fit (i.e. the fit to peoples’ perceptions of products, compared with the fit to the people that buy them). Both options have arguments in their favour. The data derived technique is more parsimonious in its generation (it does not require any extra data collection); it does not rely on subjective personification of objects; and it has the potential for further automation (it could be used to track change in personality of different brands/products over time). Therefore, this method is a time and resources efficient alternative for companies or researchers looking to assess psychological-fit within retail data. However, the Matz et al. (2016) method focuses on people’s perceptions of products, and in some settings, such as marketing, these perceptions may be more important than people’s behaviour. Furthermore, the method used by Matz et al. found match on E and A to be significant, whereas the current study’s new metric did not. This suggests there could be something different about the psychological match using people’s perceptions of a product’s A and E, compared to what people with similar A and E buy.

4.2.5 Summary

This pilot study replicated and extended the work by Matz et al. (2016). It showed that personality match to one's shopping (measured using two different methods) is related to improved Subjective Well-Being (SWB), supporting previous findings. Not only this, the analysis showed that when the psychological-match variable is broken down by trait, it is only predictive of SWB when match on traits E, A and N occurs.

A new method was proposed to calculate match using the personality of the people who bought the products, rather than using human ratings. This method, which is more parsimonious and requires less data collection, performed equally as well as the Matz et al. (2016) method on 10 product categories. It should be noted that the data derived method calls upon a slightly different conceptual interpretation of psychological match. Rather than buying products with a similar personality to oneself engendering happiness, it suggests that buying what people with a similar personality to oneself buy, makes you happier.

The novel data-derived method for calculating match has only been tested on 10 broad product categories, and many questions still remain. For example, does the same positive effect on SWB occur when personality match is calculated on a wider range of products and brands? Is psychological match associated with increased SWB for all personality types? Is the match to all products equally as good at predicting SWB, or are some products not able to capture differences in customers personality (for example if everyone buys them)? And finally, if the latter is true, can the methodology be refined to control for the fact that some products are better at capturing peoples personalities than others? The next section explores these questions, whilst conceptually linking the data-derived methodology to the methodologies harnessed by recommender systems.

4.3 Buying what people like you like: Consumer choice, personality, and well-being in big data

This study formalises the effect uncovered in Section 4.2 that buying what people like you buy predicts greater SWB. Such a phenomena is less about how a person's personality matches their shopping's and more about how their behaviour aligns with people of a similar personality, and is thus hereafter referred to as Personality Alignment (PA). That people are more likely to buy what other people of similar profiles buy, is at the core of collaborative filtering algorithms used by large companies such as Amazon and Netflix. Whilst such algorithms focus on user similarity on a behavioural level (e.g. similar purchase histories or interests), few have investigated similarity on a personality level, and the effect this can have on customer happiness.

Using the same dataset as for the pilot study, the effect of PA on individual's subjective SWB is investigated using 100 different brands and 100 product categories. In addition, this study explores whether this PA effect is contingent on the individual's personality. The

findings replicate the pilot study and show that PA on traits E, A and N positively predict customer SWB - but not O and C. Interaction effects show that the effect of PA is strongest for people high in N. It is speculated that high N individuals use shopping as a means of mood regulation. Overall, these findings provide evidence that personality and behaviour can be combined by algorithms to optimise, not just for number/type of purchases, but also for customer happiness.

4.3.1 Introduction

Recommender Systems and Customer Happiness

One way in which companies have been trying to maximise customer happiness and satisfaction has been through trying to deliver the products to a customer which are most appropriate for them as an individual. The mass amounts of data available on consumer choice has led to the rapid development of recommender systems, now an integral part of users online world (Tkalcic et al., 2009). There are three main types of recommender system: Content based recommenders (CBR), collaborative filtering recommenders (CF), and hybrid recommenders (HR) (Adomavicius and Tuzhilin, 2005; Tkalcic et al., 2009). CBRs break down a particular item into its attributes (i.e. genre/ themes/ product features) and utilise how users rate these sub-parts to build user and item profiles, on which predictions can be made (Tkalcic et al., 2009). CF systems recommend items that similar users have liked or used in the past (Adomavicius and Tuzhilin, 2005). Thus, CF is based on the presumption that when the similarity between two users is high, both users will like similar items. HRs use a combination of both CBR and CF approaches (Tkalcic et al., 2009).

As collaborative filtering systems are built on the premise that people who are alike will behave the same, this makes the similarity measure a crucial part of the CF system. The most common approach is to use behavioural similarity - people who consume similar items are rated as more similar. This assumes that people who act the same in the past will act the same in the future. Such a similarity metric incorporates preferences and interests - but these can change over time. Personality, however, is theorised to be (roughly) stable over time (Costa Jr and McCrae, 2008; John et al., 1999). The Big Five personality structure (Goldberg, 1990) which posits that people's personalities are well described by five orthogonal trait dimensions: openness to experience (O) (artistic vs. conservative), conscientiousness (C) (self-controlled vs. easy going), extraversion (E) (outgoing vs. reserved), agreeableness (A) (compassionate vs. antagonistic), and (N) neuroticism (emotionally stable vs. unstable) is widely used in consumer psychology and marketing sciences (Matz and Netzer, 2017). Recent studies have used personality to calculate the similarity between users, with promising results (Tkalcic et al., 2009; Nguyen et al., 2018).

Ultimately recommender systems want to suggest items which are going to make the user happy, as this yields greater customer satisfaction. But getting feedback on the emotional reactions to a multitude of different items is costly. One of the mechanisms that drives emotion

regulation is personality (Carver et al., 2000; Luminet et al., 2000; Davidson et al., 2001), and similar personalities predict similar emotive responses - particularly for traits E and N (Yik et al., 2002). Nguyen et al. (2018) found that different personalities showed different profiles of the diversity, popularity, and serendipity of the recommendations produced, whilst ratings-based systems failed. Tkalcic et al. (2009) found that personality based similarity measures out-performed ratings-based measures for predicting consumption. They noted the additional benefit of only a single personality measure being required (rather than many emotional ratings), yielding lower computational costs. However, whether the personality based similarity measures in CF make customers happier, is not known.

Although automatic detection of emotional affect is being developed (i.e. automatic facial expression detection during media consumption (Zeng et al., 2009)), recommender systems are rarely evaluated using customer happiness (Tkalcic et al., 2009). However, it is important to know not just that a recommendation system produced recommendations that an individual consumed, but that that consumption produced a positive emotive affect. As personality predicts emotional regulation, this suggests that CF using personality-based similarity measures could produce greater customer happiness, as well as item convergence (Tkalcic et al., 2009).

Personality Homophily and Happiness

As outlined in Section 4.1.1, the effect consumption has on improving an individual’s SWB, beyond acquiring a basic comfortable lifestyle, has remained contested for two decades, evidencing the relationship’s complex nature (Stevenson and Wolfers, 2013; Diener and Seligman, 2004; Clark et al., 2008; Di Tella and MacCulloch, 2008). While it has been shown that increased income can improve people’s SWB if money is spent in the “right” way (Dunn et al., 2011; Matz et al., 2016), contextual and psychological factors are important moderators. Personality homophily, is one of the ways in which personality may interact with happiness.

Homophily is the term used when people prefer to interact with similar others (McPherson et al., 2001). It has been found to occur for age, education, attitudes, values, and general intelligence (Rushton and Bons, 2005). Youyou et al. (2017) used behaviour and language assessments of personality (to remove the reference-group effect (Heine et al., 2002) that occurs with self-report measures) and found strong homophily effects for personality also exist. Further, Noë et al. (2018) found evidence of geographical personality homophily in social networks for trait E, but found homophilious pairs on traits O and C lived further apart than non-homophilious pairs. This suggests that homophily effects may not be uniform across all of the big Five personality traits.

If personality homophily exists it may be because it is reinforced in some way. Recent studies have looked at the effects of personality homophily on individuals’ happiness, trying to understand if the gravitation of individuals to people and situations (Sirgy (1985) which better match their personality is positively reinforced by making people happier. Jokela

et al. (2015) found clusters of people with similar personality traits living in London, and that when people had a similar level of trait O to those around them, they were happier.

Research Overview

The results of the pilot study in Section 4.2 showed that people who purchased from product categories that people with a similar personality to themselves purchased from, were happier. The notion that people are more likely to buy what other people of similar profiles buy, is at the core of collaborative filtering algorithms used by large companies such as Amazon and Netflix. Whilst such algorithms focus on user similarity on a behavioural level (e.g. similar purchase histories or interests), as noted above, few have investigated similarity on a personality level, and those that have, have not investigated whether personality similarity measures evoke greater customer happiness. Meanwhile, in the field of psychology, that personality similarity (psychological-fit) to ones shopping has been shown to make people happier (Matz et al., 2016). The present study concatenates the approach of collaborative filtering with the literature on personality homophily and psychological-fit to investigate whether people who buy items that have been bought by people of a similar personality to themselves, are happier.

This study formally introduces a new metric, personality alignment (PA), used to measure the extent to which people match the purchasing of those similar to themselves. This metric, piloted in Section 4.2, negates the need to personify inanimate objects. Instead it derives both products' and brands' "personalities" from the data, using the personality of the people who buy the product/brand - so no additional personality ratings are required. Further, this study's method requires one set of personality scores per person, compared to CF algorithms using emotive feedback requiring many affect ratings for each item consumed (Tkalcic et al., 2009). The method has also been further refined, compared to the pilot in 4.2, to incorporate the variance of people's personality traits contributing to a product/brand. Thus, this study investigates using 100 different products and 100 different brands, to explore whether using the behaviour of those who have a similar personality to oneself is an efficient and effective principle which could be utilised by CF algorithms looking to predict consumer happiness from consumption patterns.

As the literature on the effect of personality homophily on SWB is varied for different personality traits, as piloted in Section 4.2, here the work of Matz et al. (2016) is further extended by investigating for which of the traits is the effect of PA on happiness greatest. This research has applications in the development of CF algorithms using personality, as companies want to predict not just what will people buy, but what will produce greater long term happiness and satisfaction for customers.

Overall, this research extends the work of Matz et al. (2016), and that of the pilot study, in three important ways, as summarised in the inset below.

- 1) A new way to study psychological-fit is introduced, by deriving the personality ratings of items from the people who buy them - this negates having to

personify objects, and significantly reduces the amount of data needing to be collected. This methodology conceptually changes the question from ‘does buying things that match your personality make you happier?’ (Matz et al., 2016) to ‘does buying what people with a similar personality to yourself buy, make you happier?’. The latter speaks to the domain of CF which uses a similar principle of behavioural similarity, and more recently personality similarity (Tkalcic et al., 2009) to produce recommendations. The methodology is also revised from the pilot study to incorporate the variance of the customers’ personality traits who bought the particular brand/product. This gives less weighting to PA on products/brands that are bought by everyone, no matter their personality (i.e. those that have a high variance).

2) As in the pilot study, the overall personality similarity is broken down into the similarity on each of the Big Five traits separately, to analyse for *which* traits it is most important to align personality on in order for consumption to increase happiness.

3) Finally, the interactions between PA and the different traits are investigated, to explore whether PA effects SWB more for some people, with particular personality types, compared to others.

4.3.2 Method

Participants

Participants were from the sample of customers who’s completed survey responses were matched to their loyalty card data, as described in Chapter 3, Section 3.2. The first analysis used participants who had bought at least one of the most popular 100 brands ($N = 12,167$). 11,094 (91%) of the respondents were female and 1,073 (9%) were male. For the second analysis, the study used participants who had bought from one or more of the most popular 100 product categories ($N = 12,230$). 11,158 (91%) of the respondents were female and 1,072 (9%) were male. These samples are large enough to warrant generalisability of effects to the retailer’s general customer population.

Materials

The same questions for personality, well-being, and income were used as in the pilot. The brand and product categories were predefined by the retailer themselves, and recorded in the transaction data. For Analysis 1, 100 of the most popular brands were used, and for Analysis 2 the 100 most popular products. For the brands analysis, top-level brand names were leveraged (i.e. no sub-brands). For the product analysis, a finer level of product category descriptors than performed by Matz et al. (2016) was chosen. For example, where Matz et al. (2016) used catch all categories such as ‘Dental Care’, this study used the following set of product categories: Toothbrushes, Dental Floss/Tape, Toothpaste, Dental Accessories

and Mouth Fresheners. The full list of brands and product categories can be found in the Supplementary Materials (Section 10.2.2).

Analytic Approach

It was investigated whether consuming the items that people with a similar personality to oneself buy makes people happier, by analysing two different consumer behaviours: 1) what brands people buy and 2) what products people buy. This was done by formulating a metric called *personality alignment* (PA). PA represents the distance between an individual's own personality, and the expected personality of their shopping. The expected personality of the shopping can be interpreted as the predicted personality of the type of person who would buy that particular set of items. The expected personality of a person's shopping was derived from the personalities of the people who bought the products/brands in the person's shopping basket. Therefore, rather than get humans to rate the product's personalities as if they were human, the personalities of the products/brands were derived from the data - assigning a product/brand a personality using the personalities of the people who bought it.

The PA metric was calculated in steps, illustrated with the notations in *Figure 4.2*. Firstly, the *Expected Personality of Purchase* (EPP) was deduced for each product group and brand in the data. As shown in (4.2), EPP is simply a weighted mean of the trait (T) (O, C, E, A and N scores) for those people (i) who bought the product/brand (x). Each consumer's personality trait is multiplied by a weighting factor, represented as the number of times they have purchased the product ($purchase_{x,i}$), divided by the overall sales of that item ($purchase_x$). N is the total number of customers who've ever purchased the product/brand. Due to the weighted nature of the measure, if someone frequently buys a brand or product, their personality contributes more to the EPP than the personality of someone who bought it less often. As the effect of PA on the different traits separately (PA^T) was to be studied, EPP was calculated 5 times, once for each trait value (EPP^T). In summary, EPP can be interpreted as the best guess one would make, whenever a product or brand is purchased, as to the personality of its consumer.

$$EPP_x^T = \sum_{i=1}^N \left(\frac{purchase_{x,i}}{purchase_x} \times T_i \right) \quad (4.2)$$

Once the EPPs were attained, to calculate the PA metric, the difference between the individual's personality (P_i) and the EPPs of all the products/brands that person bought were calculated. This was achieved using the euclidean distance (Deza and Deza, 2009). To account for the fact that some products, such as toothpaste, would have been bought by many people regardless of their personality, the variance of the personalities that contributed to the EPP were taken into account. This was done by defining a subset (S) of traits (T) from the people who's purchases of x were greater than 0 (i.e. the people's personalities

which had contributed to the EPP for that product/brand) (see equation 4.3).

$$S_x^T = \{P_i^T | purchases_{x,i} > 0\} \quad (4.3)$$

Using this subset it was then possible to calculate the precision (τ) (or inverse of the variance) of the traits in this subset (see equation 4.4). This was to ensure that the research assigned a greater importance to the products/brands which were better at characterising a particular personality (as they had less variance).

$$\tau_x = \frac{1}{\sigma_{S_x^T}^2} \quad (4.4)$$

This precision score was then used to calculate a weighted mean of the euclidean distances between the person's personality (P) and the personality of their purchases (EPP). This was done for every product/brand a person bought (X) for each of the 5 traits, giving five different distance scores for each person (PA_i^T). It was then possible to deduct PA from the mean so that a higher PA reflected a greater alignment.

$$PA_i^T = mean - \frac{\sum_{x=1}^X (\|P_i^T - \mathcal{EPP}_x^T\|_1 * \tau_x)}{\sum_{x=1}^X \tau_x} \quad (4.5)$$

For every individual, five trait PA scores were calculated for brands (Analysis 1) and five for products (Analysis 2), reflecting the extent to which each individual's personality aligns with the expected personality of the brands/products they buy. The top 100 best selling brands and products only were used to ensure the EPP was calculated from a large enough sample to ensure a reliable estimate of the shopping's personality, and to ease computational load. Because EPPs are calculated using other shopper's personalities, if a person's PA is high (their personality highly aligns with their shopping) this indicates that they buy what people with a similar personality to themselves buy.

Predicting Spending

It was important to validate the measure of PA by testing whether people spend more on the brands and products bought by people with a similar personality to themselves. This was done by using hierarchical linear regressions with random intercepts for the product's personality. The results showed that people did spend more on products and brands purchased by people with similar personalities when demographics, and brand/product personality were

controlled. Therefore, people do buy more of what people like them buy. The full results are available in the Supplementary Materials (Section 10.2.1). In the following section the investigation into whether PA is associated with greater SWB for the individual is presented.

4.3.3 Results

Analysis 1: Brands

A sample of 12,172 individuals had bought from at least one of the 100 most popular brands, resulting in 2,589,085 individual transactions. The mean number of unique brands people bought from the top 100 was 33.68 ($SD= 15.95$), and the mean number of items people bought from each brand was 6.36 ($SD= 14.89$). See Table 4.4 for the descriptive statistics for age, average monthly spending, the Big Five, and SWB. See Section 10.2.7 of the Supplementary Materials for the correlations between the variables, and Section 10.2.3 for the personality trait density distributions.

Table 4.4: Sample Summary Statistics

	Brand Sample		Product Sample	
	Mean	<i>SD</i>	Mean	<i>SD</i>
Age (years)	51	15	51	15
Avg Monthly Spend (£)	42.77	30.37	42.74	30.36
Openness	4.84	1.19	4.84	1.18
Conscientiousness	5.84	1.08	5.84	1.08
Extraversion	4.03	1.54	4.02	1.54
Agreeableness	5.50	1.12	5.50	1.11
Neuroticism	3.31	1.46	3.31	1.46
Well-being	7.16	1.93	7.15	1.94

To investigate how PA, expressed through the brands people buy, relates to SWB, three hierarchical linear regression models were used. Model 1 was the baseline model with age, gender, income, average monthly spending, and the individual’s personality traits as control variables. Model 2 added PA on each of the Big Five. Model 3 additionally included the interaction between the person’s personality and PA on that trait. This was to additionally study whether the effect of PA on SWB was different for different personalities. The variance explained by each model, model comparisons, as well as the unstandardised coefficients and confidence intervals can be viewed in Table 4.5.

In Model 1, the baseline model with demographics, spending, and personality traits, all variables significantly predicted SWB. In Model 2, the addition of PA variables significantly

improved the prediction of SWB. Interestingly, only PA on E (PA^E), PA^A and PA^N were significant predictors of SWB. Therefore, buying brands that people with a similar level of E, A and N as you buy, predicts greater SWB. In Model 3, it was investigated whether this effect was dependent on the individual's personality traits. Model 3 showed that there was a significant interaction between PA^N and N. When the interaction effects were added, the PA^A and PA^N main effects disappeared. These interactions improved the variance explained in Model 3, but this improvement was not significant ($p=0.0569$). The interaction between N and PA^N is plotted in Figure 4.3. The interaction showed that the effect of PA on SWB was significantly greater for those people who were more neurotic.

To test the robustness of the effect of PA on SWB against additional controls, a final analysis was conducted using brands. This was to ensure that personality extremity, trait non-linearity's, and trait interactions (particularly E and N) were not responsible for the effect of PA on SWB. To do this, all the variables present in Model 3, with the addition of the Big Five's quadratic functions (Lubinski and Humphreys, 1990), the interaction between E and N (McFatter, 1994), and personality extremity (Matz et al., 2016) (the average absolute distance from the median traits score), were input into a forward and backwards processing step-wise regression predicting SWB, so that all variables were in competition with one another (Lubinski and Humphreys, 1990).

The results of the selected variables in table 4.6 showed that PA^C , PA^E , PA^A , and PA^N were selected either instead of, or as well as, the additional control variables. The effect of PA^A was reversed in the stepwise model, likely attributed to some shared variance with A^2 . N's interaction with PA^N stayed significant. Out of the control variables, the interaction E:N was selected, but O^2 , E^2 and N^2 and personality extremity were not. This analysis showed that the effect of PA on SWB was sufficiently distinct from the interaction/non-linearities of the individual's personality traits (namely E and N) acting on SWB.

Table 4.5: Hierarchical linear regressions using PA on brands to predict well-being

	<i>well-being</i>		
	Mode 1	Model 2	Model 3
Age (years)	0.01*** (0.01, 0.02)	0.01*** (0.01, 0.02)	0.01*** (0.01, 0.02)
Gender: Male	-0.32*** (-0.43, -0.21)	-0.31*** (-0.42, -0.19)	-0.31*** (-0.42, -0.19)
Income: £25,000 to £34,999	0.35*** (0.26, 0.45)	0.35*** (0.25, 0.45)	0.35*** (0.25, 0.45)
Income: £35,000 to £49,999	0.54*** (0.44, 0.64)	0.53*** (0.43, 0.63)	0.53*** (0.43, 0.63)
Income: £50,000 to £74,999	0.51*** (0.40, 0.61)	0.50*** (0.40, 0.61)	0.51*** (0.40, 0.61)
Income: £75,000 to £99,999	0.60*** (0.45, 0.74)	0.59*** (0.44, 0.74)	0.60*** (0.45, 0.74)
Income: £100,000 to £149,999	0.55*** (0.36, 0.75)	0.56*** (0.37, 0.75)	0.56*** (0.37, 0.76)
Income: £150,000 or more	0.55*** (0.27, 0.83)	0.57*** (0.29, 0.85)	0.57*** (0.29, 0.85)
Income: Prefer not to say	0.52*** (0.42, 0.61)	0.50*** (0.41, 0.59)	0.50*** (0.41, 0.60)
(log) Average Monthly Spend (£)	0.09*** (0.04, 0.14)	0.09*** (0.05, 0.14)	0.09*** (0.04, 0.14)
Openness	-0.04** (-0.07, -0.02)	-0.05** (-0.09, -0.01)	-0.05* (-0.09, -0.01)
Conscientiousness	0.13*** (0.10, 0.16)	0.13*** (0.09, 0.17)	0.13*** (0.08, 0.17)
Extraversion	0.13*** (0.11, 0.15)	0.16*** (0.14, 0.19)	0.15*** (0.12, 0.18)
Agreeableness	0.05*** (0.02, 0.08)	0.10*** (0.05, 0.14)	0.09*** (0.05, 0.14)
Neuroticism	-0.44*** (-0.46, -0.41)	-0.51*** (-0.53, -0.48)	-0.48*** (-0.51, -0.44)
PA on O		-0.02 (-0.08, 0.03)	-0.02 (-0.19, 0.16)
PA on C		-0.01 (-0.06, 0.05)	0.06 (-0.18, 0.30)
PA on E		0.08*** (0.04, 0.12)	0.16** (0.04, 0.28)
PA on A		0.07* (0.01, 0.13)	0.09 (-0.13, 0.32)
PA on N		0.17*** (0.13, 0.22)	0.03 (-0.08, 0.14)
O*PA on O			-0.001 (-0.04, 0.03)
C*PA on C			-0.01 (-0.06, 0.03)
E*PA on E			-0.02 (-0.04, 0.01)
A*PA on A			-0.004 (-0.05, 0.04)
N*PA on N			0.04** (0.01, 0.07)
Observations	12,167	12,167	12,167
R ²	0.208	0.214	0.215
Adjusted R ²	0.207	0.213	0.213
Residual Std. Error	1.72 (df = 12151)	1.71 (df = 12146)	1.71 (df = 12141)
F statistic:	$F(15,12151)=212.40^{***}$	$F(20,12146)=165.6^{***}$	$F(25,12141)=134.01^{***}$
Model Comparison		$X^2(5)=279.85^{***}$	$X^2(5)=31.501, p=0.0569$

Note: Significance levels: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. Coefficients are unstandardised b .

Table 4.6: Forwards and backwards processing step-wise regression with PA on brands predicting well-being

	<i>Dependent variable:</i>
	well-being
Age (years)	0.01*** (0.01, 0.02)
Gender: Male	-0.31*** (-0.42, -0.20)
Income: £25,000 to £34,999	0.35*** (0.25, 0.45)
Income: £35,000 to £49,999	0.53*** (0.43, 0.63)
Income: £50,000 to £74,999	0.50*** (0.40, 0.61)
Income: £75,000 to £99,999	0.59*** (0.44, 0.74)
Income: £100,000 to £149,999	0.56*** (0.37, 0.75)
Income: £150,000 or more	0.57*** (0.29, 0.85)
Income: Prefer not to say	0.50*** (0.41, 0.60)
(log) Average Monthly Spend (£)	0.09*** (0.04, 0.14)
Openness	-0.04** (-0.07, -0.01)
Conscientiousness	-0.50 (-1.13, 0.13)
Extraversion	0.11*** (0.05, 0.16)
Agreeableness	0.16 (-0.31, 0.63)
Neuroticism	0.17 (-0.02, 0.36)
PA ^C	0.07*** (0.03, 0.11)
PA ^E	0.10** (0.03, 0.16)
PA ^A	-0.38** (-0.66, -0.10)
PA ^N	0.07* (0.0000, 0.13)
C ²	0.01*** (0.01, 0.02)
A ²	-0.08** (-0.13, -0.03)
N*PA ^N	0.09*** (0.05, 0.13)
E*N	0.02* (0.002, 0.03)
Observations	12,167
R ²	0.216
Adjusted R ²	0.215
Residual Std. Error	1.71 (df = 12143)
F Statistic	$F(23,12143)=145.42^{***}$

Note: Coefficients are unstandardised *b*. Significance levels are * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.
The variables that were not selected were: Personality Extremity, PA^O, O², E², N², O*PA^O, C*PA^C, E*PA^E, and A*PA^A

Analysis 2: Products

Slightly more individuals (12,230) had bought from at least one of the 100 most popular product categories than brands, resulting in 3,188,290 individual transactions. The mean number of product categories people bought from was 41.35 ($SD=17.11$), and the mean

number of products people bought in each category was 6.39 ($SD=14.39$). The descriptive statistics for age, average monthly spending, the Big Five, and SWB can be found in Table 4.4. See Section 10.2.7 of the Supplementary Materials for the correlations between the variables, and Section 10.2.3 for the personality trait density distributions.

To investigate how PA, expressed through the products people buy, relates to SWB, three hierarchical linear regression models were used - using the same method of analysis as was done for brands. The variance explained by each model, the model comparisons, and the unstandardised coefficients and confidence intervals can be viewed in Table 4.8. In Model 1, with demographics, spending, and personality traits only, all variables significantly predicted SWB. In Model 2, the addition of PA on the Big Five significantly improved the model. As with brands, only PA^E , PA^A , and PA^N were significant predictors of SWB. Therefore, buying brands that people with a similar level of E, A and N as you buy, predicted greater SWB. In Model 3, it was investigated whether this effect was dependent on the individual's personality traits. Model 3 showed that there was a significant interaction between PA^N and trait N. This interaction significantly improved the variance explained in Model 3 compared to Model 2. When interaction effects were included, the main effects of PA^A and PA^N disappeared. The interactions are plotted in Figure 4.4. The interaction plot shows that PA on products, the same as with brands, has a significantly larger positive effect on well-being for people who are more neurotic.

To test the robustness of the effect of PA on SWB against additional controls a final analysis using products was conducted, as it had been with brands. To ensure personality extremity, trait non-linearity's and interactions were not responsible for the effect of PA on SWB, all variables present in Model 3, and the additional controls, were input into a forward and backwards processing step-wise regression. The results of the selected variables are shown in Table 4.9. The step-wise regression showed that PA^E and PA^A remained strong positive predictors of SWB, but PA^N 's effect was reversed (likely due to the inclusion of N^2 in the model). The interaction of $N:PA^N$ remained a strong significant predictor. The $E:N$ interaction was significant, but personality extremity was not selected, neither was O^2 , E^2 , or A^2 .

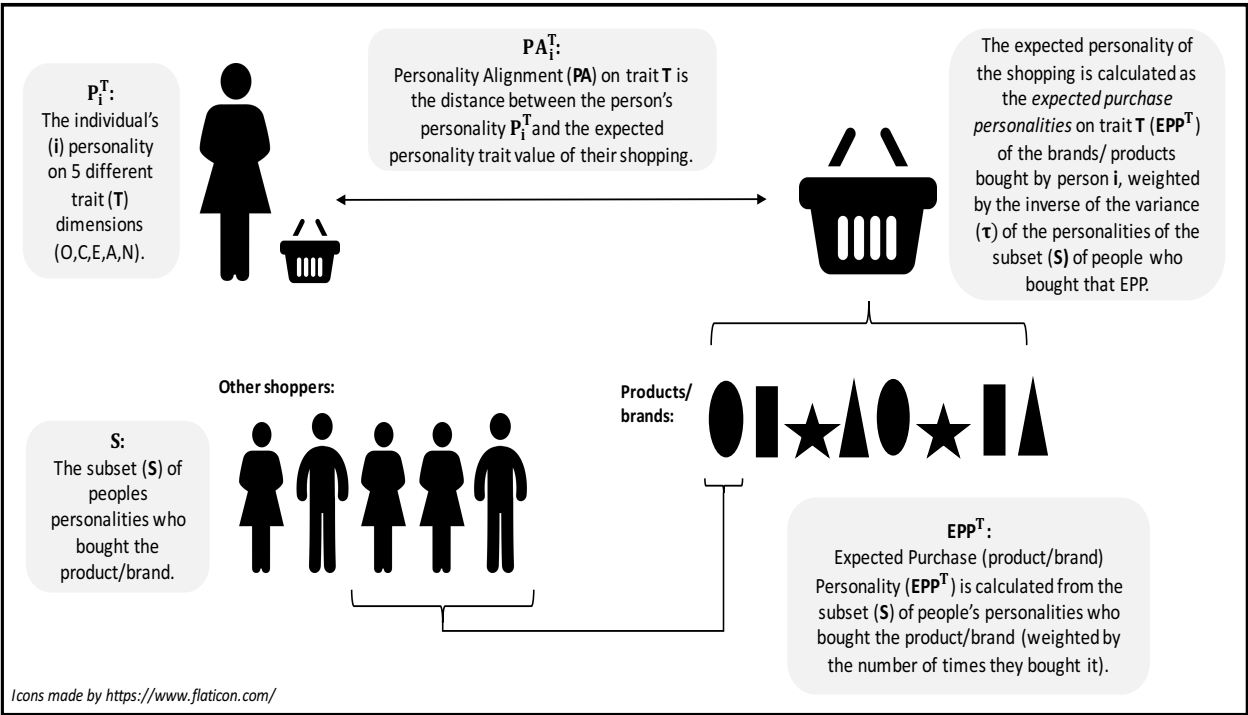


Figure 4.2: Illustration of the method used to calculate personality alignment (PA)

Table 4.7: Exemplar product categories strongly emphasized by particular personalities (all at least 1.5 SD from the mean for the trait)

+O	Salads, Fresh Fruit, Baby Toiletries
−O	Hand Wash, Fragrance Gifts, Body Sprays
+C	Contact Lens, Cotton Wool, First Aid
−C	Jewellery, Sweet Snacks, Dietary Products
+E	Self Tan Products, Contact Lens, Tights
−E	Sanitary Towels, Dietary Products, Incontinence Pads
+A	Decongestant, First Aid Accessories, Denture Products
−A	Baby Toiletries, Nappies, Formula/Milk
+N	Sanitary Towels, Sweet Snacks, Newspapers/Magazines
−N	Smoking Cessation, Contact Lens, Decongestant

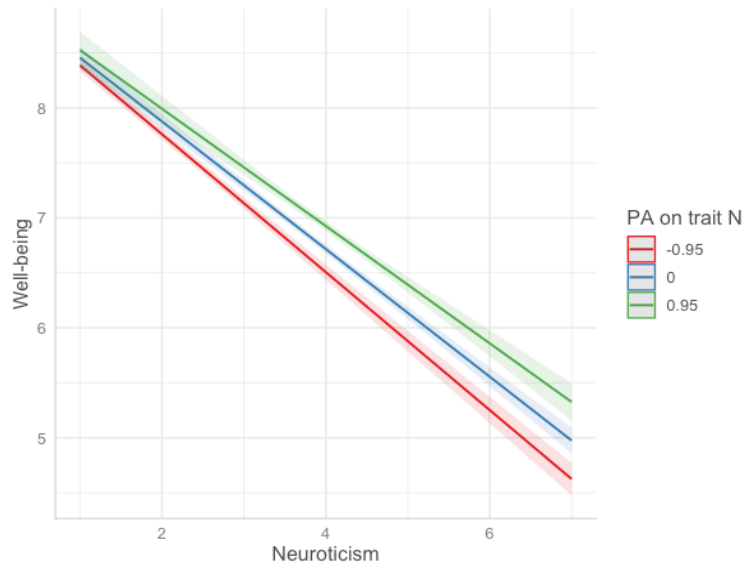


Figure 4.3: A graph to show the interaction between PA^N for brands and Neuroticism on well-being

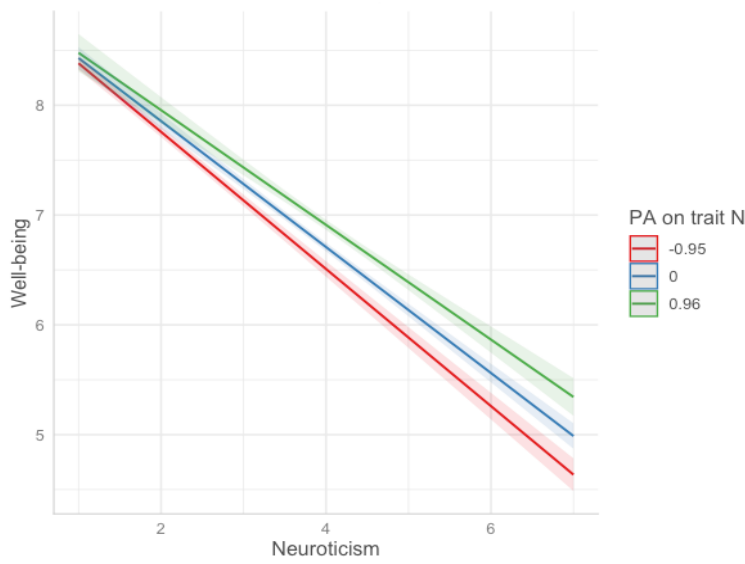


Figure 4.4: A graph to show the interaction between PA^N for products and Neuroticism on well-being

Expected Personality of Purchase and Variance

The analyses threw up some clear differences between the ‘brand/product personalities’ assigned to items in previous work, and the personality type of consumer that was actually found to be buying such products in reality. Matz et al. (2016)’s Dental Care category was rated by Amazon Turks as being more than -1 SD in O and +1 SD in C. However, this analysis found that the categories Toothpaste and Toothbrushes appeared entirely neutral in terms of EPP - i.e. they can only ‘indicate’ a consumer of average personality type. The combined Z -scores across the Big Five were only 0.27 for Toothpaste, and 0.21 for Toothbrushes. This is compared to a product category like Smoking Cessation where the combined Z -score across the Big Five was 3.44.

This demonstrates the difference between enforcing personifications of brands/products and measuring the personalities of the people who buy them. However, this is representative

of the fact that some products are simply poor indicators of personality. Some products are ‘neutral’ because their consumers have a greater variance in personalities (products with a low variance are much better at indicating a particular personality profile). This makes sense for necessities, items which can only ever be neutral given that people have to buy them irrespective of their own personality. Neutral products, however, give a yardstick, and this work uses the product category Toothbrushes as a neutral, uninformative category against which can be used to compare the traits of other product categories’ E, A, and N profiles. Figure 4.5 shows four examples where product categories vary in insightful ways:- Smoking Cessation (purchased by those higher in extraversion and emotional stability); Self Tanning products (higher in extraversion); and Incontinence Pads (introversion and agreeableness) are illustrative examples which are considered in more detail below (a more detailed list of EPPs for all products and brands used in this study is supplied in the Supplementary Materials (Section 10.2.2, and exemplar product categories with high EPPs are listed for each trait in Table 4.7).

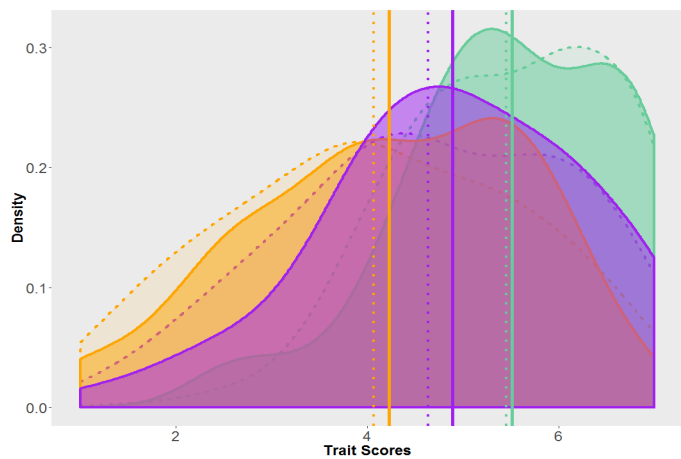
The product category Smoking Cessation, was purchased by individual’s with higher E ($t(28411)=10.21$, $p<0.001$, $d=0.10$), lower neuroticism (N) ($t(28470)=17.32$, $p<0.001$, $d=0.18$), and slightly higher A ($t(26226)=5.13$, $p<0.001$, $d=0.05$) than average. Those quitting smoking (and getting help from such products) therefore appear more likely to be agreeable and emotionally stable extroverts. Research has shown that smokers are on average more neurotic and less agreeable than non smokers (Terracciano and Costa Jr, 2004), but this analysis shows that such personalities are also neglecting smoking cessation help. This category’s expected personality also demonstrates the misconceptions which may arise when rating a product category. For example, people may connect Smoking Cessation to smokers (after all only smokers purchase such items), branding it as neurotic, when actually the category is for those smokers who are trying to give up - both agreeable and emotionally stable. As demonstrated, this is an important distinction to make, as the two sub-groups (smokers and smokers trying to give up) have opposing levels of A and N.

Three other examples are presented here of how a product can indicate the personality of the people who buy it. Using the significance criteria of $p<0.001$, Self Tan products were bought by people who had average levels of N (their personality did not significantly differ from those who bought toothbrushes) ($t(22382)=1.69$, $p=0.092$, $d=0.02$) and average A ($t(21748)=1.98$, $p=0.048$, $d=0.02$) - yet they had much higher than average levels of E ($t(23176)=13.58$, $p<0.001$, $d=0.15$). Thus, it can be inferred here that self tanning products are not only more likely to be bought by those who enjoy going out and socialising, but also reliable signals for such a personality.

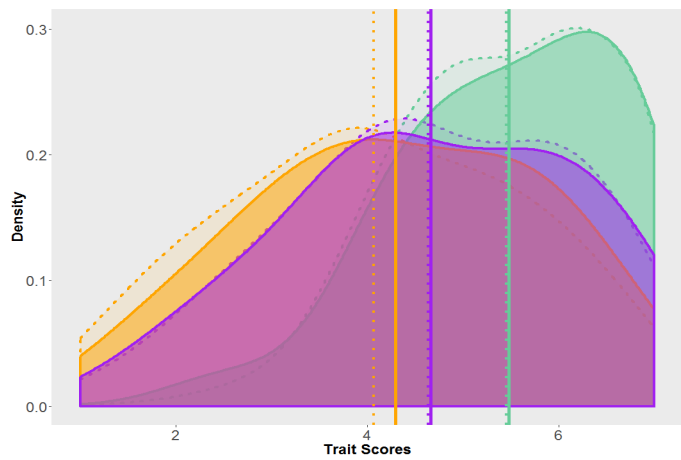
Baby Milk/Formula were bought by people who were average in E ($t(22936)=-2.26$, $p=0.024$, $d=0.03$), but less emotionally stable ($t(23809)=-18.82$, $p<0.001$, $d=0.20$) and less agreeable ($t(23632)=-18.95$, $p<0.001$, $d=0.20$). This may be indicative of the traits of young parents/mothers in general, or may highlight psychological differences between mothers who breastfeed and mothers who don’t. Incontinence Pads were slightly higher in Emotional

Stability (ES) ($t(22385)=3.42$, $p<0.001$, $d=0.04$), higher in A ($t(23399)=22.38$, $p<0.001$, $d=0.25$), but lower in E ($t(22682)=-9.10$, $p<0.001$, $d=0.10$) which is hypothesised to indicate the potential social anxiety that people suffering with incontinence may experience, which in turn can render them more introverted.

In summary, while it has only been touched upon the potential insights made available through examining EPP, the graphs in Figure 4.5 demonstrate the distinction that the methodology reveals between uninformative products (with high variance and an average or ‘neutral’ EPP) and informative ones (strong indicators of a particular personality profile due to consumers with specific traits being more likely to buy them). Moreover, it has been shown that these ‘indicated’ personalities can differ to how people would personify them if asked to imagine a personality for them.



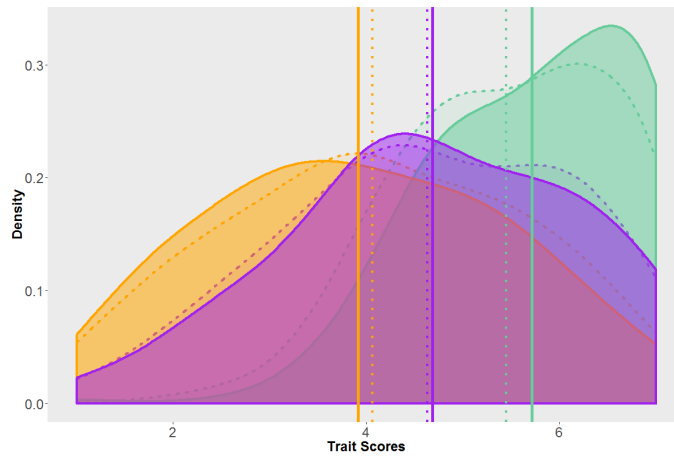
(a) Smoking Cessation



(b) Self Tan Products



(c) Baby Milk/Formula



(d) Incontinence Pads

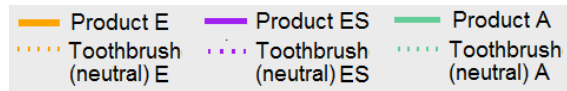


Figure 4.5: Four examples of the distribution of personalities buying certain product categories. If a person had bought a product X amount of times their personality was added X times before calculating the means and distributions of the products' EPP. Extraversion (E) is in orange; Emotional Stability (ES) (the reverse of N, where the greater the value the more emotionally stable) is in purple; agreeableness (A) is in green. The more opaque coloured areas and hard lines indicate each product's distribution compared with the faded areas and dotted lines representing the neutral category Toothbrushes distribution. The strong vertical lines show the mean trait values for the product, and the vertical dotted lines show the mean trait values for Toothbrushes.

4.3.4 Discussion

The analyses show that personality alignment (PA), as the similarity between an individual's self-reported personality and their personality reflected by their consumption, positively predicts SWB. For both brands and products, this effect was greater than the effect of average monthly spending. This provides further evidence that it is *what* you spend your money on that matters for SWB (Dunn et al., 2011; Matz et al., 2016). It was also found that PA^E , PA^A and PA^N only (not PA^O and PA^C) predicted SWB, and that the effect of PA^N

Table 4.8: Hierarchical linear regressions using PA on products to predict well-being

	<i>well-being</i>		
	Mode 1	Model 2	Model 3
Age (years)	0.01*** (0.01, 0.02)	0.01*** (0.01, 0.02)	0.01*** (0.01, 0.02)
Gender: Male	-0.32*** (-0.43, -0.21)	-0.30*** (-0.42, -0.19)	-0.30*** (-0.42, -0.19)
Income: £25,000 to £34,999	0.37*** (0.27, 0.47)	0.36*** (0.27, 0.46)	0.37*** (0.27, 0.47)
Income: £35,000 to £49,999	0.55*** (0.45, 0.65)	0.54*** (0.44, 0.64)	0.55*** (0.45, 0.65)
Income: £50,000 to £74,999	0.51*** (0.41, 0.62)	0.51*** (0.41, 0.62)	0.52*** (0.41, 0.63)
Income: £75,000 to £99,999	0.61*** (0.46, 0.75)	0.60*** (0.45, 0.75)	0.61*** (0.46, 0.75)
Income: £100,000 to £149,999	0.56*** (0.36, 0.75)	0.57*** (0.38, 0.76)	0.57*** (0.38, 0.77)
Income: £150,000 or more	0.56*** (0.27, 0.84)	0.58*** (0.30, 0.86)	0.57*** (0.29, 0.85)
Income: Prefer not to say	0.50*** (0.40, 0.59)	0.48*** (0.39, 0.58)	0.49*** (0.39, 0.58)
(log) Average Monthly Spend (£)	0.08** (0.03, 0.12)	0.08*** (0.03, 0.13)	0.08** (0.03, 0.13)
Openness	-0.04** (-0.07, -0.01)	-0.05** (-0.09, -0.02)	-0.05* (-0.09, -0.01)
Conscientiousness	0.13*** (0.10, 0.16)	0.13*** (0.09, 0.17)	0.12*** (0.08, 0.17)
Extraversion	0.14*** (0.11, 0.16)	0.17*** (0.14, 0.20)	0.16*** (0.13, 0.19)
Agreeableness	0.04** (0.01, 0.07)	0.09*** (0.05, 0.14)	0.10*** (0.05, 0.14)
Neuroticism	-0.44*** (-0.46, -0.41)	-0.50*** (-0.53, -0.47)	-0.47*** (-0.51, -0.44)
PA ^O		-0.04 (-0.09, 0.01)	-0.05 (-0.23, 0.12)
PA ^C		-0.01 (-0.07, 0.05)	0.11 (-0.13, 0.35)
PA ^E		0.09*** (0.05, 0.13)	0.14* (0.02, 0.27)
PA ^A		0.08** (0.02, 0.14)	0.05 (-0.18, 0.27)
PA ^N		0.16*** (0.12, 0.20)	0.002 (-0.11, 0.11)
O*PA ^O			0.003 (-0.03, 0.04)
C*PA ^C			-0.02 (-0.07, 0.02)
E*PA ^E			-0.01 (-0.04, 0.01)
A*PA ^A			0.01 (-0.04, 0.05)
N*PA ^N			0.05** (0.02, 0.08)
Observations	12,230	12,230	12,230
R ²	0.21	0.21	0.21
Adjusted R ²	0.20	0.21	0.21
Residual Std. Error	1.73 (df = 12214)	1.72 (df = 12209)	1.72 (df = 12204)
F Statistic	F(15,12214)=210.38***	F(20,12209)=163.69***	F(25,12204)=131.53***
Model Comparison		X ² (5)=281.29***	X ² (5)=36.566*

Note: *p<0.05; **p<0.01; ***p<0.001

Table 4.9: Forwards and backwards processing step-wise regression with PA on products predicting well-being

	<i>Well-being</i>
Age (years)	0.01*** (0.01, 0.02)
Gender: Male	-0.31*** (-0.42, -0.19)
Income: £25,000 to £34,999	0.36*** (0.27, 0.46)
Income: £35,000 to £49,999	0.54*** (0.44, 0.64)
Income: £50,000 to £74,999	0.51*** (0.40, 0.62)
Income: £75,000 to £99,999	0.60*** (0.45, 0.74)
Income: £100,000 to £149,999	0.57*** (0.37, 0.76)
Income: £150,000 or more	0.58*** (0.29, 0.86)
Income: Prefer not to say	0.49*** (0.39, 0.58)
(log) Average Monthly Spend (£)	0.08** (0.03, 0.13)
Openness	-0.05** (-0.09, -0.02)
Extraversion	0.11*** (0.05, 0.17)
Agreeableness	0.09*** (0.05, 0.14)
Neuroticism	0.30 (-0.17, 0.78)
PA ^O	-0.04 (-0.09, 0.01)
PA ^E	0.08*** (0.04, 0.12)
PA ^A	0.08** (0.02, 0.14)
PA ^N	-0.48*** (-0.77, -0.20)
C ²	0.01*** (0.01, 0.02)
N ²	-0.10*** (-0.15, -0.04)
N*PA ^N	0.10*** (0.06, 0.14)
E*N	0.02* (0.002, 0.03)
Observations	12,230
R ²	0.213
Adjusted R ²	0.212
Residual Std. Error	1.72 (df = 12207)
F Statistic	$F(22,12207)=150.48^{***}$

Note: Coefficients are unstandardised *b*. Significance levels are **p* <0.05; ***p* <0.01; ****p* <0.001. The variables that were not selected were: Conscientiousness, Personality Extremity, PA^C, O², E², A², O*PA^O, C*PA^C, E*PA^E, and A*PA^A

on SWB was greater for people who were more neurotic. Because of how PA was calculated, using the personality and behaviour of others, the analysis showed that consuming items that people with a similar personality to oneself buy, is related to increased happiness. Whilst previous approaches have either been focused on behavioural similarity with other users (as with CF algorithms), or personality similarity with products (Matz et al., 2016), this study's findings suggest that the two can be meaningfully integrated to drive, not only higher conversion rates, but also higher levels of customer subjective well-being.

The work by Matz et al. (2016) was extended here to look at the alignment on each of the big five separately - rather than combined together as one score. For both brands and products it was found that only personality similarity on E, N and A were important traits to align on for SWB. Such information is useful for CF algorithms, because if O and C are truly redundant they would not need to be measured, reducing the information required at the start to calculate personality similarity. The findings of this work are somewhat aligned with Noë et al. (2018) who found a geographic personality homophilious effect for E, and a negative effect for O and C. The findings however contradict Jokela et al. (2015) who found that a geographic homophilious effect for O was related to increased SWB. This highlights that the personality similarity effects for different traits on SWB may be context dependent, or depend on how personality alignment is calculated.

In Model 3 for both brands and products, when the interaction between PA and the person's personality was taken into account, the effect of PA^A and PA^N went away completely. The interactions show that the effect PA has on SWB is much stronger for neurotic people than emotionally stable people. Yet PA is equally related to SWB for both introverts and extroverts. This suggests that introverts are happiest when they buy things that other introverts buy, and extroverts are happiest when they buy things that other extroverts buy. The stronger effect of PA on SWB for high N people could be because people with greater mood instability could be using consumption to cheer themselves up more. High N products include Sweet Snacks, Jewellery, and Newspapers/Magazines - all of which could be used to temporarily alleviate negative affect. Thus, a person high in N who buys these products (the products that other people who are high in N use to cheer themselves up) would be happier. Further research could use experience sampling methods to study the effect of purchases on mood, and purchase-induced moods on SWB over time. Overall, these interactions show that the effect of personality alignment in a consumer environment is more complex than was first reported by Matz et al. (2016).

The current work looked at PA on 100 brands as well as 100 product categories, extending Matz et al. (2016). Both brand and product alignment were significant predictors of SWB, with similar effect sizes, despite reflecting very different item groupings. This suggests it is the alignment, not which items are bought, that is important to SWB. The results, however, are correlational - it remains a statistical challenge to understand whether happier people buy in line with their personality, or whether behaving in line with your personality itself engenders increased SWB (or both).

The calculation of PA presented in this study is a computationally efficient way of calculating ‘psychological-fit’ as it negates the need to gain trait ratings of products and brands to attribute a, somewhat arbitrary, personality score to them (Avis et al., 2014). Instead, just one initial set of ratings is needed to gain customers personality scores - and the above analysis shows that the TIPI is sufficient for this. Ratings are usually required for CF algorithms anyway to alleviate the ‘cold start’ problem, where initially no behavioural data is available to calculate similarity. Future methods, however, could look at extrapolating customers personality from their shopping (Gladstone et al., 2019), and incorporating this into recommender systems to alleviate the annoyance of customers having to carry out a personality questionnaire altogether.

Behavioural indicators of personality revealed in Big Data is a fruitful area of research which has mostly been constrained to online behaviour (Back et al., 2010; Quercia et al., 2011) until recently (Gladstone et al., 2019). This work shows that certain products, with tighter variance, can reveal specific personality types. For example, Contact Lens purchases are a strong indicator of conscientious and extroverted personalities; using Dietary products is a consistent expression of neurotic, introverted personalities; and so on. The method, used here, of calculating EPP holds potential to reveal further insights into the psychological make-up of those making particular consumer choices.

As well as the large quantity that it is possible to collect, a further advantage of using transaction data is that it constitutes a measure of real world behaviour, unbiased by social desirability (Van de Mortel et al., 2008) or misremembering (Jakovcevic et al., 2014), as self-reported behaviour measures can be. A limitation of working with such data, however, is that it is assumed the individual’s personality is related to every purchase they buy. However, it is to be expected that people are also buying items for other people - something which cannot be controlled for. Although the literature on personality homophily suggests that friends and partners have similar personalities (Youyou et al., 2017), and these people are likely to be the ones that customers are buying for, if not for themselves, this will still introduce noise into the data - contributing to smaller effect sizes being found.

This study did not engineer a recommender system, but the effect uncovered - that people who consume items that people with a similar personality to themselves consume, are happier - is a principle that practically could now be harnessed by CF algorithms. Whilst CF traditionally optimises for purchases, these results show that systems built using PA could be optimised for both purchases (see results in Supplementary Materials, Section 10.2.1) and customer SWB. Compared to the measure used by Matz et al. (2016) for psychological-fit, this study’s data-derived PA technique is more efficient (it does not require extra data collection); is less idiosyncratic (as noted above, it does not rely on the personification of objects); it has the potential for further automation (EPP can be automatically updated live as different people with different personalities buy different things, and could even be calculated on an individual item level.) In addition, the effect sizes that this study found, were substantially larger than those found by Matz et al. (2016) (which have been publicly

contested (Boyce et al., 2017)). It is worth noting, however, that the method used by Matz et al. (2016) focuses on people's perceptions of products, and in some settings, such as marketing, these perceptions may be more important than people's actual behaviour. Overall, this study provides compelling evidence that personality and behaviour can be combined by algorithms, not just to optimise for purchases, but also for happiness.

Chapter 5

Predicting Well-being in Big Data using Machine Learning Methods

5.1 Introduction

The previous chapter built on research by Matz et al. (2016) who found that the match between a person's personality and what they buy predicts greater subjective well-being. A year after Matz et al. published their work, Boyce et al. (2017) refuted the methodology used and the subsequent conclusions drawn. They demonstrated, with simulations, that in big data small random effects can be found significant. Although they did not have access to the same bank records data, they randomly generated profiles for participant's 'shopping baskets' from participants personality data in the English Longitudinal Study of Ageing (ELSA) to investigate the likelihood of detecting a spurious effect when one would not be expected, on a similar size to that used by Matz et al. (2016) ($N=625$). They repeated the randomisation 1000 times, and found that greater basket-person match, when randomly generated, positively predicted life satisfaction ($\beta=0.12$). This occurred in 80.3% of their replications (where $p < 0.05$) and when they used the full ELSA sample ($N=7,990$) it occurred in 100% of the simulations.

Such an analysis does not disprove the results from Matz et al. (2016), but it does call into question the legitimacy of using p -values to detect effects in big data when even random effects can easily be significant. Boyce et al. (2017) also highlight that even if the effect were not resultant from randomness, it is so small (particularly compared to the effect personality, for example) that it is not meaningful. The study presented in Section 4.3 found greater effect sizes than Matz et al. (2016), yet the sample size used was greater than theirs, creating a greater likelihood of mistaking noise for a statistically significant effect. The study additionally controls for quadratic and interaction effects, which Matz et al. (2016) did not, strengthening the conclusion that it is the *alignment* that is responsible for the effect, not some other artefact of the shopping's personality, or the person's personality (or their combination).

The results of Matz et al. (2016)'s study suggest that the extent to which an individual's

spending matches their personality explains less than 1% of the variance in Subjective Well-Being (SWB). Matz et al. (2017) argue that, unlike personality, most people have a degree of choice over what consumables they spend their money on. Thus, it is important to draw particular attention to the small things people can *change* to improve their happiness, as opposed to the big things they can not. Furthermore, small effects can add up. An example of this is the Abelson’s Paradox (Abelson, 1985) in baseball, where small cumulative effects in batting performance culminate to large differences between batting wages over a season. Likewise, small increases in well-being on an individual level, could have large social effects.

Consumption is a fundamental part of people’s life. Even the very poorest groups in the world (i.e. those living on less than \$2 per day) spend considerable amounts on discretionary goods, such as entertainment, celebrations, clothing and tobacco (Banerjee and Duflo, 2007). If recommender systems built by consumer giants like Amazon were tuned to encourage people to buy things that will make them happier, rather than to just buy anything, tiny increments in people’s mood can occur. If just the tiniest increase in happiness can be yielded from each purchase a person makes, over the course of a week, month, or year, gains could theoretically accumulate rapidly for an individual. When this is scaled to the 244 million Amazon customers (Kline, 2014), a small effect can generate meaningful social impact.

Boyce et al. (2017) argue, however, that it is misleading to persuade the public that consumption is the key to happiness, particularly as materialism has been shown to negatively effect well-being (Dittmar et al., 2014). But in the case of subscriptions to something like Netflix, or where consumers are already searching Amazon looking for something they have already decided to buy, then providing customers with recommendations which are going to optimise customer satisfaction, and even consumer well-being, can only be seen as positive (squeezing the maximum happiness from one’s money).

It is a well documented phenomena that even minuscule effects can be significant if the sample is large enough. Having more data available makes more complex models, with more variables and more parameters, seem more appealing. Yet if these models are just fitted to the data, inflated effect sizes can occur (Yarkoni and Westfall, 2017). Machine learning models are complex algorithms that perform well on big data, by learning on one set of data, and predicting on another. Therefore, when tested on unseen data, they uncover only effects which generalise to new samples. This can help to prevent overfitting and reduce the likelihood of false discovery (Yarkoni and Westfall, 2017). The next section investigates the differences between fitting algorithms to data versus predicting out-of-sample. It also introduces decision trees as a way to find out more about the non-linear structures, and potential subgroups in the data (which can be more easily detected in larger datasets).

5.2 Using Machine Learning to Gain Additional Insights

The following work outlines an extension to the previous Chapter, Section 4.3, ‘Buying what people like you like: Consumer choice, personality, and well-being in big data’. It first investigates how machine learning methods, such as cross-validation, can help to mitigate the issue of over fitting in large datasets, and the potential misidentification of noise as a true or meaningful effect, as discussed above. This section illustrates the difference between out-of-sample prediction and the variance explained metric commonly used in psychology, and how the former can help determine small but reliable effects from randomness.

Section 4.3 found that Personality Alignment (PA), demographics, and personality traits linearly predict SWB. An additional step-wise regression analysis studied the quadratic effects of traits Extraversion (E) and Neuroticism (N), and their interaction. The analysis was driven by previous literature about the relationship between traits E, N, and SWB. But in some cases there isn’t past research available to guide decisions on whether or not to include interaction or quadratic effects in a regression. This is the case with the newly engineered variable PA. It is problematic to hypothesise about how it may interact with different demographic variables, for example, when there is no literature to make these assumptions from. Further, even if a quadratic effect isn’t significant, some other non-linear relationship could still exist, one which doesn’t fit a particular mathematical function. As discussed in Chapter 2, decision trees are useful for exploring non-linearities, interactions, and subgroups in the data, using the print out tree structure. In this section, it is illustrated how decision trees and random forests can offer an alternative method to the step-wise regression in Section 4.3 for identifying non-linear predictor variables.

5.2.1 Insight 1: Fitting Models can lead to Inflated Effect Sizes

The larger the dataset, the greater the complexity of the models or algorithms that can be used on it. However, the more complex models are, the more likely they are to over fit. As discussed in Chapter 2, this can be mitigated using methods of model evaluation such as testing on a hold out set, or cross validation. To investigate the difference out-of-sample testing makes for models of different complexities, five models were evaluated using different evaluation metrics. The same data and variables were used as in Section 4.3 Analysis 2, Model 3 (not inclusive of the interaction or quadratic variables) using PA on products to predict SWB. The Multiple Linear Regression’s (MLR) performance was compared to a baseline (mean prediction), a Decision Tree (DT), Random Forest (RF), and a Neural Network (NN). The models were evaluated using Cross Validation (CV) (see Section 2.1.4) Root Mean Squared Error (RMSE), CV Mean Absolute Error (MAE), R^2 on training data, and CV R^2 .

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \tag{5.1}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - x_i)^2}{n}} \tag{5.2}$$

Table 5.1: Model Comparison

Model	CV RMSE	CV MAE	R^2 Training Data	R^2 CV
Baseline (Mean)	2.34	1.82	-	-
Multiple Linear Regression	1.72	1.31	0.21	0.21
Decision Tree	1.75	1.33	0.89	0.18
Random Forest	1.72	1.31	0.82	0.21
Neural Network	1.74	1.32	0.28	0.19

Results

The different models’ performance, evaluated using different metrics, can be compared in Table 5.1. Both the RMSE and the MAE measure the average magnitude of the errors in a set of predictions (independent of the direction). The MAE is the mean value of the absolute differences between the actual and the predicted values (see Equation 5.1). The RMSE is the square root of the average of the squared differences between the actual and predicted values (See Equation 5.2). Both metrics range between zero and infinity, where the closer to zero the better (less error). The RMSE penalises large errors more heavily. For example, if a prediction error of 4 is more than twice as bad as a prediction error of 2, the RMSE is the more appropriate metric to use. It is not hypothesised that this is the case for PA and well-being, and Table 5.1 shows that the type of error metric used does not change the rankings of the models. The R^2 on the training data, measures the variance explained when the model is fitted to the training data. The CV R^2 is the average variance explained across all 10 test folds in the CV procedure.

All the models produced a similar CV MAE, and sufficiently less than the mean baseline prediction. The largest difference in performance is when evaluating the models using the R^2 on the training data. Using this metric the decision tree, random forest, and neural network perform substantially better than the linear regression. However, when comparing the models using the CV R^2 , the models perform similarly. Note that the performance evaluation of the linear model, does not change between using the two different R^2 metrics, indicating it generalises well and is not over fitting to the training data.

The CV MLR performance is illustrated in Figure 5.1. The graph shows the regression lines for all 10 folds layered on top of each other making them impossible to distinguish. This means the model varies very little across different sub-samples of the data (it is highly replicable). This is expected given it is a relatively simple linear model on a large sample size. However, the linear relationship is very generalised, as Figure 5.1 illustrates, there is a lot of data not being captured by the linear model.

Discussion

The model comparisons illustrate how, even on large datasets, complex models can overfit to data (here the training sample). Importantly, on small sample sizes a model doesn't need to be very complex to overfit (Yarkoni and Westfall, 2017), is thought to be a serious issue in psychological research where small to moderate effect sizes are expected (Yarkoni and Westfall, 2017). That there was no change in the value for the R^2 compared with the CV R^2 for the MLR shows that large sample sizes can mitigate against overfitting.

This analysis showed that the MLR performed just as well as the more complex RF algorithm. This could indicate three different possibilities 1) that there are mostly linear relationships between variables in the data, which the RF is modelling using smooth decision boundaries; 2) that both models reached the total amount of generalisable variance explained within the data (the rest is noise, and modelling this would lead to overfitting); or 3) both models represent two equally good but different solutions. Differentiating between these theories is difficult in this analysis, as the machine learning models' meta-parameters were not optimised for out-of-sample prediction. This is typically done in a validation stage (using cross validation on the training data) to select the parameters which ensure the model generalises best to new data. This would likely improve the more complex models' performance, like the NN, relative to the simple regression. In the next chapter, this model optimisation is undergone to compare how different models fare when they are tested on new data.

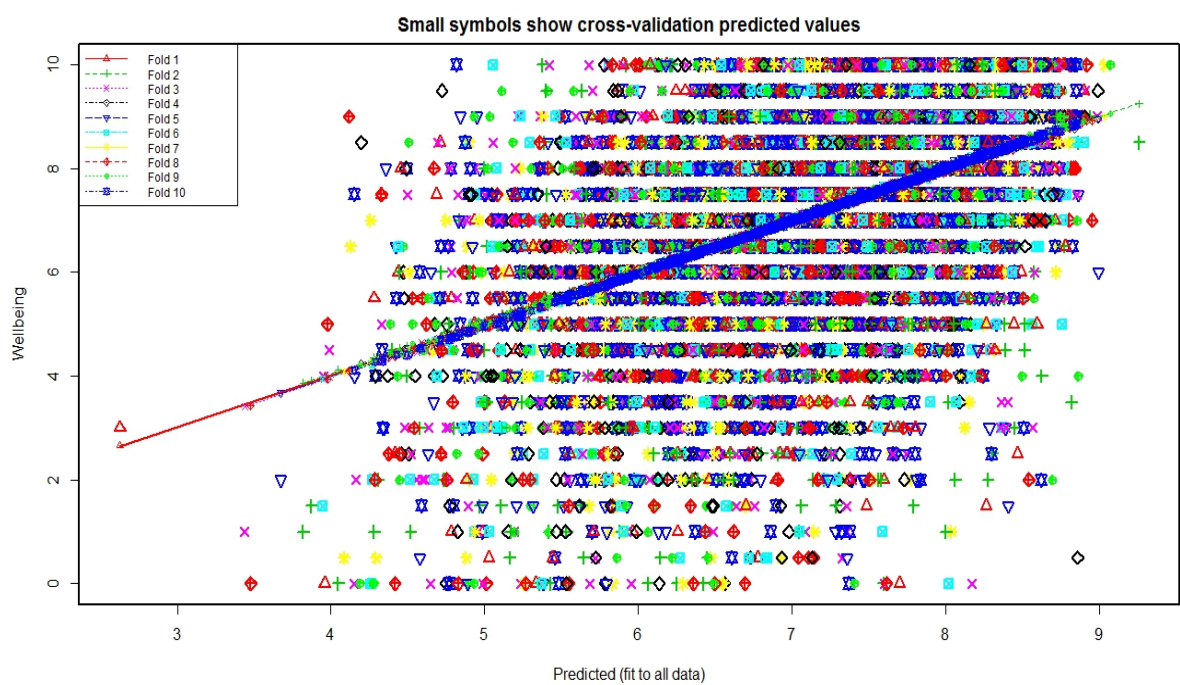


Figure 5.1: A graph to show the fit of the linear regression on all 10 folds of the cross validation. The x-axis shows the predicted well-being scores, and the y-axis shows the actual well-being scores.

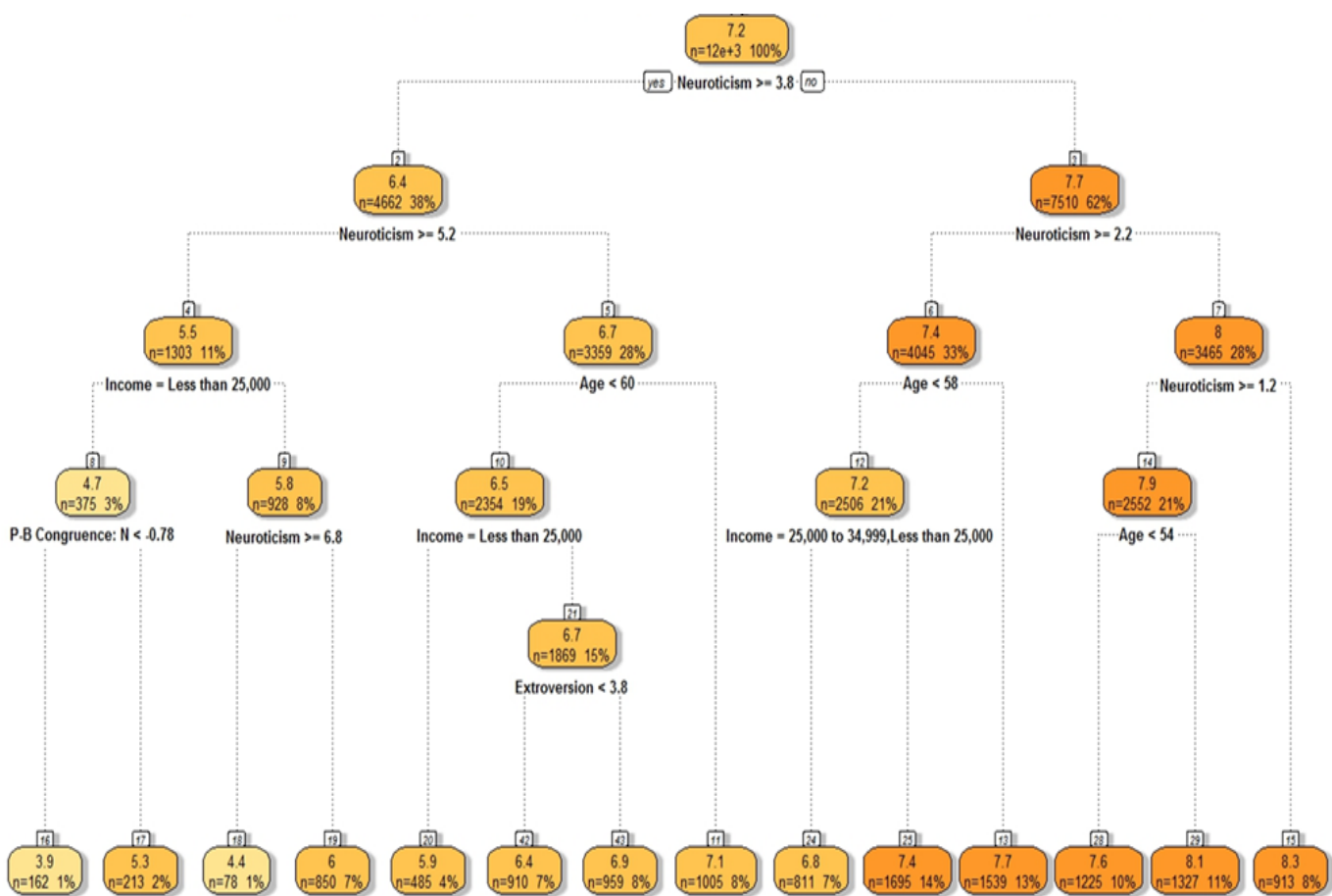


Figure 5.2: A decision tree to predict well-being scores (0-10) using customer's personality traits, age, gender, income, average monthly spending, and PA for brands. The decision tree was pruned at the point in which cross-validation was lowest.

5.2.2 Insight 2: Decision Tree's can be used to Gain Additional Understanding

The model comparison's in the previous section showed that the DT did not perform as well as the MLR using CV. However, the DT was within 3% variance explained by the MLR using much less information (just 13 binary decisions). This shows that identifying only the most important parts of a variable can reduce the information the model requires. The variables and the points on the variables that were useful can be studied in the tree structure to reveal more about the factors predicting well-being.

The decision tree structure can be viewed in Figure 5.2. The tree gives more insights into the contingencies and interactions between the variables. From looking at the tree structure, N is used in the top three splits, indicating that it is the most important for predicting SWB. Once the tree splits the data into four groups using N, other variables emerge as being important, and these are different for different sub-groups of N. These variables are age, income, PA^N , and E.

The first node splits the data on N at the point of greater or equal to 3.8/7. The mean N is 3.3, and therefore a slightly higher than average N has a detrimental effect on well-being. N can be used to predict the people with the highest SWB (node 15) (score of 8.3) with no input from any other variables. Thus this demonstrates a simple relationship between N and well-being where those who are highly emotionally stable are very happy. On the other (left) side of the tree the reverse simple relationship does not exist. Other variables in combination with N are used to predict low SWB. For example, at node 4 having a low income (when people are already above average in N) predicts the difference between 4.7/10 and 5.8/10 SWB. Interestingly, a low PA^N predicts the people who are most unhappy.

By studying the left (low SWB) side of the DT structure, variables can be identified which have a protective effect on the negative relationship between N and Well-being. For example, if a person has a disposition to poor well-being because they are higher than 3.8 in N, this can be counteracted (to varying degrees) by having greater than £25,000 annual income, being older than 60 years, being more extrovert (i.e. having a greater access to social support), and buying things that are consistent with their level of N.

On the right (high SWB) side of the tree are the factors which have a detrimental effect on SWB for emotionally stable individuals. If people are not exceptionally low in N, younger than their mid-late 50's, and do not have an income greater than £34,999 a year, they are predicted a SWB score of 6.8/10 (node 24), lower than the average SWB score of 7.1. This is also lower than the people on the left hand side of the tree who are higher in N, but are older than 60 (node 11); or have a greater income than £25,000 a year, and are more extrovert (node 43). Overall, this decision tree demonstrates the intricate interactions which exist between personality traits and situational factors when predicting SWB.

The MLR used all 14 variables, whilst the tree only uses 5. Further, the DT only used two of the categories on the income variable. Therefore, when using the decision tree to predict

consumer well-being much less information about customers' is required. In this example, this comes with a price of a 3% decrease in accuracy. Yet, using a DT more information about the variable interactions and the important points on the variables has been gained. Such insights can inform a reduction in future data collection. Furthermore, by partitioning the data, the decision tree shows that for one group of people (i.e. high N, low income) the variable PA^N is highly predictive of SWB, but not necessarily across other parts of the population. This illustrates how in situations where it is costly to target everyone (i.e. for an intervention) decision trees can help find the areas/subgroups which will be most susceptible to a particular treatment.

Variable Importance Comparisons

A key difference between the multiple linear regression (MLR) and the DT is that the latter is non-linear. Therefore, the tree's variable importance measures reflect the variables' non-linear prediction power. By comparing the variable importance measure for the tree and the MLR model, one can compare each variable's linear and non-linear (including their inclusion in subgroups and interactions) prediction power.

To illustrate this, PA^N has the 6th largest effect size in the linear model (see Figure 5.3), however in the variable importance graphs for both the tree (Figure 5.4) and the RF (Figure 5.5) it is ranked as the 2nd best predictor. This suggests that the relationship of PA^N and SWB is non-linear, or is particularly strong for certain subgroups in the data (as illustrated in the DT structure in Figure 5.2). Further investigation showed a weak linear correlation between PA^N and SWB ($r = 0.04$). The relationship with a non-linear smoothing using a generalized additive model (GAM) can be found in Figure 5.6. The graph illustrates that there is a strong positive linear relationship with SWB for the lower half of the PA^N variable, then after a value of approximately -0.02 it levels off. When a 2nd order (quadratic) polynomial of PA^N was entered into a simple regression predicting SWB, PA^N was significant ($t(12169) = -13.7, p < 0.001$).

A major difference between the DT and the RF variable importance rankings is that income is ranked 3rd by the DT, yet 9th by the RF. In the RF the PA variables are all ranked higher than income, and the person's personality traits - whilst in the DT the PA variables are less important. In general, as the random forest is the more accurate model, its feature importance rankings have greater validity than the trees. Thus, it is likely that PA has important non-linear effects on SWB. However, having solely a measure of incremental node purity, without being able to view the tree structure, entails that the interpretation of the RF features is limited. Therefore, using both DT and RFs in combination with one another allows for a useful strategy for detecting and studying the non-linearity in the data.

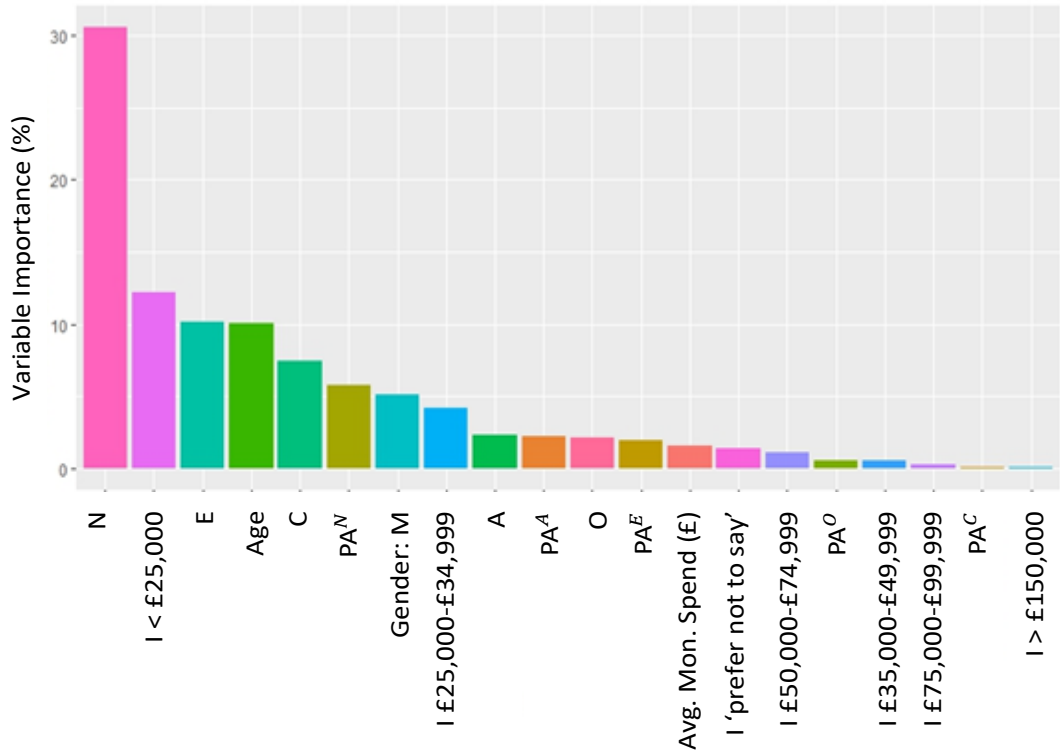


Figure 5.3: A variable importance graph showing the variable importance as a percentage (calculated using standardised β coefficients) of the variables in the linear model predicting well-being

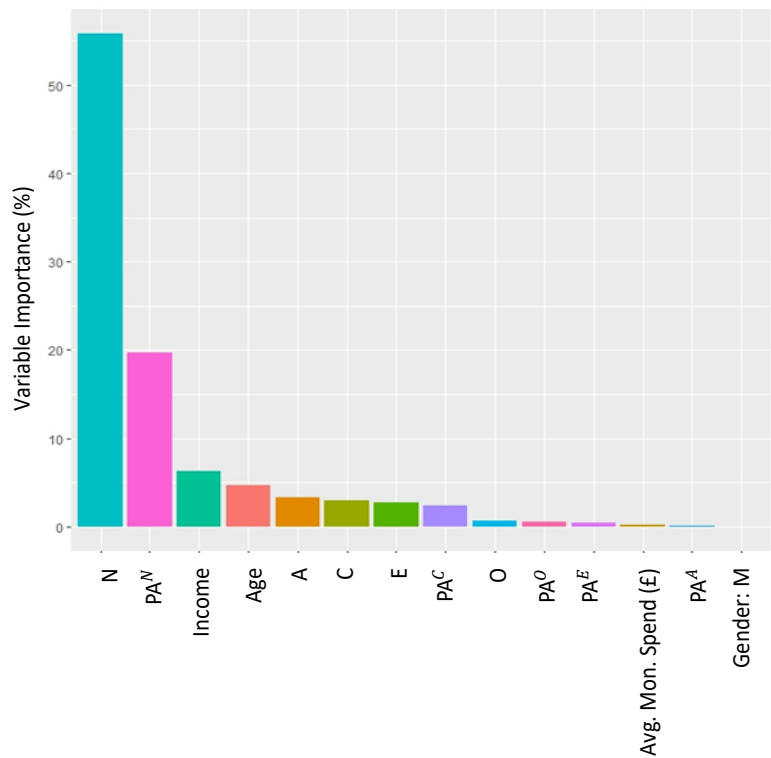


Figure 5.4: The decision tree's variable importance graph showing the relative importance (as a percentage) of the variables predicting well-being. Variable importance is calculated as the total reduction of node impurities across the whole of the pre-pruned tree.

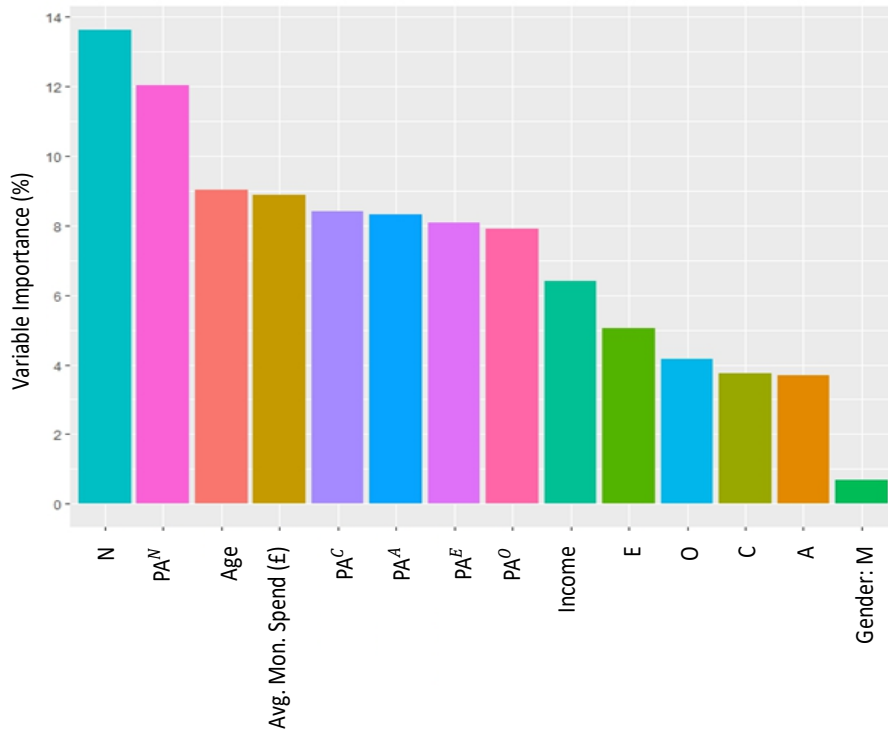


Figure 5.5: The random forest’s variable importance graph showing the relative importance (as a percentage) of the variables predicting well-being. Variable importance is calculated as the average reduction of node impurities across all of the 500 trees in the forest.

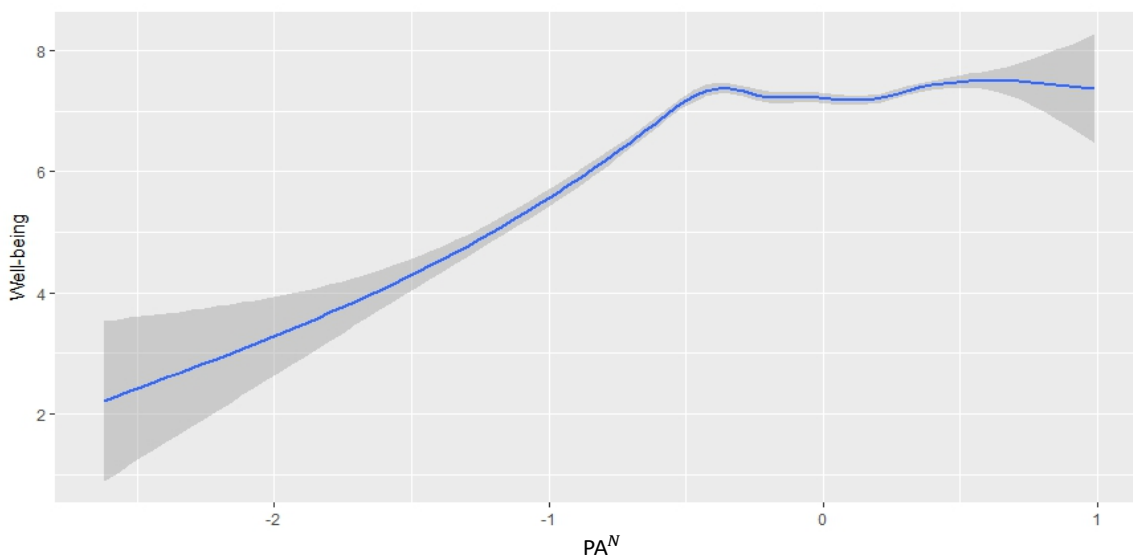


Figure 5.6: A graph to show the non-linear relationship of PA^N and well-being using a generalised additive model (GAM) fitted to the data.

5.3 Summary

This section has outlined a case study on how a machine learning approach, in particular using decision trees, can enhance psychological investigation. Firstly, the evaluation of the models using an explanation approach (fitting to training data) versus a prediction approach (using CV) highlighted the problem of inflated R^2 values. This emphasised the necessary caution that should be applied when fitting more complex models such as DTs, RFs and

NNs to data - emphasising why, in later chapters, I employ approaches such as CV and out-of-sample testing. It was also demonstrated how decision trees can be a useful tool for teasing apart how features may interact when predicting SWB, and how they can prompt linearity or subgroup investigations. Overall, decision trees and random forests can be useful additions to big data research, aiding both interpretation and prediction.

Chapter 6

Understanding the Predictors of Plastic Bag Consumption in Big Data

In the previous chapter, Subjective Well-Being (SWB) was first predicted using a linear regression model. This model was pre-selected prior to knowing much about the structure of the data, and chosen primarily on the basis of previous literature (Matz et al., 2016). This is the approach taken predominately in psychology and the field of statistics (Breiman et al., 2001). However, with more datasets becoming accessible for research, the *data modelling* approach (Breiman et al., 2001) where a model is selected based on assumptions, is not the only viable option anymore.

Section 5.2 illustrated how decision trees and random forests can be used to gain additional insights into the factors predicting SWB. The Cross Validated (CV) R^2 vales were very similar to that of the original linear model - and for the Random Forest (RF) it was the same. However, what if a tool which is highly interpretable, performs significantly worse than a more complex model which doesn't lend itself well to interpretation? Or, alternatively, what if two models perform the same, yet disagree on what variables are important?

These questions, are among some of those raised by an *algorithmic modelling* approach, where multiple models are tried and evaluated before selecting one. Such an approach is utilised in this chapter to classify people who frequently buy plastic bags at Boots. This chapter aims to use both prediction and explanation to better understand the psychological and demographic predictors of plastic bag buying.

This chapter commences with literature on plastic bag use, environmental behaviours and personality traits. It then introduces the advantages of algorithmic modelling prior to starting the study's aims and predictions. Analysis 1 uses an algorithmic inductive approach to select the model that performed the best. Analysis 2 refines the research question by engineering the Dependent Variable (DV) differently, and focuses on how best to extract the *statistically* significant predictors.

6.1 Introduction

6.1.1 Plastic Bag Charges

In 2014, 7.6 billion plastic bags were used by customers of English supermarkets (GOV, 2015; WRAP, 2015). That is on average 140 bags per person (GOV, 2015), and equates to approximately 61,000 tonnes of plastic (Smithers, 2016). Plastic bags are typically made from non-biodegradable high density polyethylene (HDPE), a material sourced from fossil fuels (Edwards and Fry, 2011). As well as the energy and oil used to create them, most bags are not recycled (Monaghan, 2018) and negatively impact the environment when littered or inadequately disposed of (Jambeck et al., 2015). Plastic bags contribute to the 12.7 million tonnes of plastic that end up in our seas each year (Jambeck et al., 2015), polluting the waters and harming marine wildlife (Jambeck et al., 2015). Many national and local governments have introduced policies to try to reduce the number of bags used (Clapp and Swanston, 2009), these include bans (e.g. in Kenya (NEMA, 2018)), and mandatory charges (e.g. in Ireland in 2002 (Anastasio and Nix, 2016)). In October 2015, England followed Scotland, Wales, and Northern Ireland in introducing a mandatory 5 pence charge for single use plastic bags, in an attempt to reduce the number of single-use plastic being consumed.

In England, the 5 pence charge has been successful in reducing the number of bags used by shoppers (DEFRA, 2018b) mirroring findings from other countries (Welsh Government, 2016; Poortinga et al., 2013; Scottish Government, 2015; Block, 2013; Haoran, 2010; Australia, 2015). England’s six-month review post-levy showed that the number of single-use plastic bags had dropped by more than 85% to around 500 million in six months (DEFRA, 2018b). People’s attitudes towards the levy have also changed (Poortinga et al., 2016; Thomas et al., 2019). A longitudinal survey found that six months after the levy, there was increased support for the charges, and improved awareness of the environmental impact of household plastic waste (Poortinga et al., 2016; Thomas et al., 2019). When people reduced their plastic bag consumption, it was for both financial and environmental reasons (Jakovcevic et al., 2014; Poortinga et al., 2016). A supermarket observational study showed that there were no changes in the number of customers bringing personal bags, such as rucksacks, to the supermarket. The use of ‘bags for life’, however, doubled after the levy – suggesting shoppers were replacing single-use bags with bags for life (Poortinga et al., 2016).

Studying real world environmental behaviour unobtrusively, whilst retaining detailed information about the individual, is a difficult task. Diary interview studies have teased apart people’s motivations and attitudes for buying plastic bags (Poortinga et al., 2016), but when using smaller samples it is hard to attest generalisability. Questionnaire and longitudinal surveys have reached a wider participant pool (Poortinga et al., 2016; He, 2010; Sharp et al., 2010), but self-reported behaviour measures can be affected by social desirability bias (Van de Mortel et al., 2008) or misremembering (Jakovcevic et al., 2014). Kormos and Gifford (2014) in a meta-analysis found that 79% of the variance in the association between self-reported pro-environmental behaviour and objective pro-environmental behaviour remains

unexplained. This is especially troubling when intervention campaigns are being devised on the results of studies using self-reported behaviour. Therefore, if actual behaviour can be measured accurately on a large scale, along with people's psychological attributes and demographic information, more accurate insights into who and what drives plastic bag buying in the real world can be made, and used to inform future policy decisions.

Demographic Predictors of Environmental Concern

Understanding the profile of people who frequently buy plastic bags, in spite of the levy, is important for planning future plastic reduction interventions, legislation, and campaigns. Previous research has shown that people concerned with the environment are more likely to be younger, female, have more money, and higher education than people less concerned with the environment (Davidson and Freudenburg, 1996; Hines et al., 1987; Kalof et al., 2002; Zelezny et al., 2000). Unsurprisingly, pro-environmental individuals hold positive environmental attitudes (Dunlap et al., 2000), are well informed on the impact of their behaviour (Frick et al., 2004), and demonstrate concern about the natural world (Bamberg and Möser, 2007). However, whether this general profile of pro-environmental individuals is the same for specifically those who avoid purchasing single-use plastic bags by bringing their own bags to the store, has not yet been fully studied.

Observers in Jakovcevic et al. (2014) and Poortinga et al.'s (2016) supermarket observational study, estimated bag buyers' age and gender. The latter found that in England the use of single-use plastic bags decreased from 53% to 23% among men, and from 55% to 20% among women after the levy. Shoppers estimated to be 18-44 years bought the most plastic bags, whilst older people (estimated to be 65 years or over) bought the least - despite research suggesting older people are less concerned about the environment (Gifford and Nilsson, 2014). This supports literature finding that younger adults are more concerned with the environment, but older adults engage in more pro-environmental behaviour (Arcury and Christianson, 1993; Honnold, 1984; Klineberg et al., 1998). This suggests that certain actions such as plastic bag buying are less driven by environmental concern and instead by other factors - perhaps frugality (Tapia-Fonllem et al., 2013), anti-consumerism (Jackson, 2012), or being cautious with money (Poortinga et al., 2016; Jakovcevic et al., 2014)?

Personality and Environmental Concern

Research has demonstrated that personality can influence people's environmental concern. Early studies found that pro-environmental attitudes were associated with being more liberal (Arbuthnot, 1977), and conscientious (Pettus and Giles, 1987). Recent research has focused on using the Big Five framework (Costa and MacCrae, 1992) of measuring personality. Milfont and Sibley (2012) showed that on both an individual level and national level, Agreeableness (A) (sympathetic, warm), Conscientiousness (C) (dependable, self-disciplined) and Openness to New Experiences (O) (intellectual, complex) positively predicted environmental engagement. A British study found that agreeableness and conscientiousness were positively

related to recycling behaviours (Swami et al., 2011). Hirsh and Dolderman (2007) and Hirsh (2010) found high A and O strongly, as well as Neuroticism (N) (anxious, easily upset) and C weakly, positively related to environmental concern. Neuroticism (N) was additionally shown by Wiseman and Bogner (2003) to positively relate to certain aspects of environmental engagement. Evidence for the role of Extraversion (E) (fun, enthusiastic), however, has been mixed (Borden and Francis (1978); Milfont and Sibley (2012); Hirsh (2010)).

Markowitz et al. (2012) and Fraj and Martinez (2006) found that O consistently related to pro-environmental activities in both a community sample and an undergraduate student sample. They suggested that high levels of aesthetic appreciation, creativity, and inquisitiveness, rather than altruism, may motivate pro-environmental actions. People high on trait O have a greater willingness for change, to break with convention, and experience new ways of doing things (Markowitz et al., 2012). High O individuals are also driven by aesthetic appreciation and intellectual curiosity - both of which can influence one's interest in nature and environmentalism (Hirsh and Dolderman, 2007; Hartig et al., 2001). In summary, multiple studies have found that different personality traits, namely O and A, but also C and N positively relate to environmental concerns and engagement. However, the extent to which these relationships translate from general pro-environmental attitudes and activities to the specific behaviour of purchasing plastic bags is unknown.

Engaging in environmental behaviour involves not just the motivation or intention to do so, but additionally the self-discipline and belief in the behaviour's impact (Van der Linden et al., 2015; Axon, 2017) in order to carry them out. A greater internal locus of control (the degree to which people believe that they have control over the outcome of events in their lives) has been associated with stronger pro-environmental intentions and behaviour in both Germany and Japan (Ando et al., 2010), Australia (Fielding and Head, 2012) and Canada (Abrahamse et al., 2009). Thus when studying real world behaviour, it is not just the predictors of beliefs, attitudes and concerns that are important - but also the dispositional traits which affect whether an action is carried out or not. These factors may include: self-control, impulsivity, future focus, propensity to plan ahead, and self-efficacy - none of which are explicitly measured using the Big Five.

Future orientation (or future time preference), the tendency to establish goals and to plan strategies for meeting long-term obligations (Corral-Verdugo et al., 2006) is consistently positively related to sustainable behaviours (Corral-Verdugo et al., 2006; Joireman et al., 2004; Urien and Kilbourne, 2011). Frugality refers to being economic with resources, and antagonist to consumerism (Jackson, 2012) and materialism (Goldsmith and Flynn, 2015). Consumerism is one of the essential causes of environmental problems (Tapia-Fonllem et al., 2013). As well as reduced consumption, frugality affects the type of items purchased, and the methods used for waste disposal. Unsurprisingly, studies show that frugality relates to pro-environmental behaviour (Tapia-Fonllem et al., 2013; Gatersleben et al., 2014; Pepper et al., 2011). As well as impacting through self-transcendent values such as unity with nature (Todd and Lawson, 2003), there is also a self-enhancement aspect of frugality which relates to

pro-environmental behaviour - ambition and self-control, as well as consumer independence (Goldsmith and Flynn, 2015). Low impulsivity (akin to high self-control) relates to the ability to plan ahead captured in time preferences, and is in part captured in trait O (people high in O are often impulsive (McCrae, 1993)) however there is limited literature explicitly on the inhibition of impulses and pro-environmental behaviour.

Personality and Plastic Bag Purchasing

Personality can predict environmental concerns, engagement and self-reported behaviour. However, how this literature translates specifically to bag buying behaviour in the real world is uncertain for three reasons: 1) The variables influencing one environmental behaviour measure, such as smart meter-readings, may not generalise to another, i.e. plastic bag purchasing. 2) The studies which measure the effect of personality on environmental concern can not necessarily be generalised to actual environmental actions and behaviour. There are many potential mediating and moderating variables that can influence concerns translating to actions. 3) Finally, studies which measure self-reported behaviour rely on participants being able to remember past behaviour accurately, and report behaviour without bias. The following inset explores these points more fully.

1) One environmental behaviour does not always predict another. There is evidence of environmental behaviour spillover from plastic bag reduction, where one ecological behaviour can lead to another (Thomas et al., 2016; Lanzini and Thøgersen, 2014), but it does not always occur (Poortinga et al., 2013; Thøgersen and Ölander, 2003), depends on the type of behaviour (i.e. the cost (Lanzini and Thøgersen, 2014)), and in some cases one behaviour could even reduce the propensity to behave environmentally friendly in other areas (Thøgersen and Ölander, 2003). Ego-depletion theory (Baumeister et al., 1998) proposes people have a finite amount of will-power or self control. Once this is used up, people are unable to exert the same level of self-control on subsequent, often unrelated tasks or choices. For example: “I spent ages sorting the recycling this morning - I didn’t have time to think about bringing a bag for life with me”. Thus the factors predicting one environmental behaviour (recycling household waste) may not be predictive of another (i.e. bringing a personal shopping bag to the store) - particularly if one behaviour is habitual and the other is not.

2) Limitations exist on the extent to which environmental attitudes and intentions generalise to pro-environmental behaviour. For example, Diamantopoulos et al. (2003) showed that socio-demographic variables could define consumer profiles for environmental knowledge and attitudes, but were limited for actual behaviour. Environmental knowledge and values have been found to account for only 40% of the variance in ecological behavioural intentions, which in turn explains just 75% of the variance in ecological behaviour (Kaiser et al., 1999). Thus, it is uncertain whether the same personality traits which influence envi-

ronmental concerns in the literature, predict actual pro-environmental behaviour - specifically the purchasing of plastic bags.

3) Studies which measure self-reported behaviour using questionnaires may be affected by bias (Van de Mortel et al., 2008) or poor memory (Jakovcevic et al., 2014), and may not truly represent behaviour occurring in the real world (Kormos and Gifford, 2014). Kormos and Gifford (2014) found that in the association between self-reported pro-environmental behaviour and objective behavioural measurements, the majority of the variance was unexplained. This can lead researchers to draw misleading conclusions about the usefulness of theories or predictive models tested on self-reported behaviour.

Environmental Behaviour in Big Data

Until recently, objective environmental behaviour has been hard to measure on a large scale, as it has typically involved observing people in the real world without them knowing (Poortinga et al., 2016; Kormos and Gifford, 2014), gaining peer ratings (which can be equally biased/inaccurate) (Chao and Lam, 2011), or looking at household meter readings (which are often not at the level of the individual) (Fuj et al., 1985). Field studies allow researchers to study a behaviour from a snapshot in time at a particular location. They cannot measure the *regularity* or patterns in an individual’s behaviour. Today, people leave a substantial amount of data in their wake. Smart meters installed into homes, smart appliances, mobile phone records, apps, and transaction records are all sources of data which can now be used to study real world ecological behaviour unobtrusively. Using loyalty cards and personal accounts, researchers can now track and record individual’s environmental behavioural patterns through time and space. With permission, this behaviour can additionally be paired self-reported psychological information. Such a unique dataset is exploited to study the psychological and demographic predictors of frequent plastic bag consumption *en mass*.

6.1.2 Algorithm Modelling versus Data Modelling

As big data psychology highlights cultural differences between psychology and machine learning in their approach to research, a similar process is occurring in the field of statistics (Secchi, 2018; Faraway and Augustin, 2018; Dunson, 2018). Statistics is the science of collecting, analysing, and interpreting data, whilst accounting for the relevant uncertainties. In big data research, statistics can help with the quality and missingness of the data, assessing causality with observational data, detecting and controlling for bias in the data, and quantifying uncertainty around predictions.

Breiman et al. (2001) highlights the two cultures that emerge when the goal is to use statistical models to reach conclusions from data: data modelling and algorithmic modelling. The data modelling culture, also referred to as the ‘classical statistics approach’, assumes that data is generated by a stochastic process (along a random probability distribution), and

requires us to choose a model using knowledge of the structure of the data (Bzdok et al., 2018). Inferences are then made based on the model selected. The algorithmic culture, encompassing machine learning methods, treats the data generation process as unknown, complex, and at least partly unknowable. The algorithmic modelling culture requires us to choose a model based on its empirical capabilities (Bzdok et al., 2018). The two cultures are not completely independent. Methods such as bootstrapping cross over both approaches (e.g. being used for statistical inference and forming the basis of many ensemble machine learning algorithms). Like psychology, the field of statistics primarily takes the data modelling approach (Breiman et al., 2001) (estimated 98%), with the field of algorithmic modelling for the most part being developed by computer scientists (Breiman et al., 2001).

Classical statistics (data modelling) were designed for long (rather than wide) data where there are only a number of independent variables and ‘sample sizes that would be considered small to moderate today’ (Bzdok et al., 2018). As the number of variables increase in a model, the number of associations and possible interactions also increase. Thus the model increases in complexity to capture these relations, and the boundary between when statistical and ML approaches are appropriate becomes blurred (Bzdok et al., 2018). As raised in Chapter 3, big data research in the social sciences predominantly still uses parametric models, despite parametric models being unlikely to fully capture the variability and complexity in big data.

Additionally, with many features, models fitted to big data can overfit, and the accuracy of the model becomes inflated (Breiman et al., 2001). This has been shown to occur for goodness of fit tests (Bickel et al., 2006) and residuals analyses (Breiman et al., 2001; Cleveland, 1993). Thus, by imposing data modelling on big data, not only do we risk losing information, but we are likely inflating the model’s accuracy.

Model Selection

When using data modelling, many researchers will fit a model i.e. a linear regression without any consideration of whether it is the best fit for the data (i.e. whether the data could have been generated by a linear model). By forcing a model to fit the data, conclusions are made based on the model’s mechanism, not the underlying patterns in the data. Thus, despite the model being a statistically significant fit to the data, the data could in fact be being generated by a non-linear and non-normal distribution process. Because the assumptions being made at the model selection stage are wrong, the interpretation of the model is also flawed. As the data size increases and the data patterns become more complex, applying the correct fitting model to the data first time is unlikely. Hence, when dealing with big data it is important to evaluate each different model’s fit.

For any given research question multiple models can be statistically significant, and these models can give diverging interpretations (McCullagh and Nelder, 1989). That multiple significant prediction models may have different interpretations when large amounts of natural data exists fits with what we know about human behaviour in the real world. As well as noise from unmeasured/unmeasurable variables, there are multiple causal pathways and interac-

tions amounting to a person performing a complex action. It is unrealistic to expect a single model to explain an individual's behaviour let alone a population's. For example, in Chapter 5, a particular level of well-being is reached (or not) by many interacting, overlapping causal mechanisms, shown using a decision tree model. Therefore, evaluating and selecting the best model can be crucial. Predicting out-of-sample and evaluating each different model's prediction accuracy can help with this process. Out-of-sample testing or cross validation helps to inform researchers of which models may have been overfitting by evaluating them empirically.

Different models, such as random forests and linear/logistic regression models, project through the data space in different ways. Thus, if one performs better than the other, this tells us something about the structure of the data. If a random forest out performs a logistic regression model, then it can be inferred that there is heterogeneity in the data that the linear regression cannot model. For example, there may be non-linear subgroups or interactions between the variables. Therefore, when the two models perform the same, there is likely to be one of 3 scenarios occurring: 1) Both models have come to the same solution through different means (using different combinations of predictor variables, but misclassifying the same cases); 2) Both models have reached different solutions which are equally as good (have used different predictor variables and misclassified different cases); or 3) the more complex model is emulating the solution that the simpler logistic regression model made but in more 'moves' (using the same predictors and classifying the same cases).

Model selection is rarely analysed in psychology. With the majority of research using smaller samples, and the trend towards pre-registering studies, model selection is supposed to be decided prior to the collection of data, based on structural assumptions of the foreseen data. However, in big data research the data structure is complex and mostly unknown. Thus, applying models where the data structure needs to be inferred, without alternative models being tested, seems unreasonable. Because of this, Analysis 1 uses an algorithmic approach to select the most accurate model from which to extract feature importance measures.

Simplicity versus Accuracy

An advantage of using a data modelling approach is that it produces a simple and understandable representation of the relationship between inputs and the output, if that the model does fit the data. When predicting, accuracy and simplicity (aiding interpretation and use) often conflict each other. Much of the focus is on accuracy, but simplicity is not unimportant. Being able to understand a model is crucial for both policy changes and interventions to take place. For example, imagine a case where a model could predict an incurable disease accurately. Although valuable for insurance companies, the model is of little use to physicians trying to cure the disease. The physicians need to know how the model works in order to look for the cure or prevention. That the disease can be predicted in itself provides physicians with some information, as it indicates that there is some detectable pattern that a

machine can identify and model. However, funding and policy decisions are made by humans not machines. If highly qualified doctors cannot interpret what the predictors are and how they are related, how will they be able to disseminate the knowledge, and communicate the results to potential funders for further research?

Simplicity is useful. But ultimately researchers want to make correct insights or conclusions from the data - coming to any old conclusion won't do. Thus, it is better to start with the most accurate model possible and try to interpret it, than to easily understand an ill-fitted model and extract a wrong conclusion. Breiman et al. (2001) states that 'interpretability' shouldn't be the goal of statistical modelling - instead 'accurate information' should. A model doesn't have to be simple to provide reliable information about the relationship between inputs and outputs. Therefore, choosing a model primarily for its simplicity is not advised, unless 1) there is reason to believe the process being modelled is simple, or 2) only simple models can be of any use to the application.

In this research, the models being selected are constrained to include only somewhat interpretable models, with the most complex being a random forest. This is because the goal of this research is *understanding* the predictors of regular plastic bag buying. This involves both prediction and explanation. And if a model is not interpretable, it is limited in how much it can contribute to our understanding of the problem. Thus, the following section presents an inductive algorithmic investigation into plastic bag consumption, using algorithmic modelling, variable importance, and post-hoc analyses.

6.2 Psychological and Demographic Predictors of Plastic Bag Consumption in Big Data

6.2.1 Aim

Despite the success of the levy, there are still around a billion single-use plastic bags bought each year in England alone (DEFRA, 2018b). In this study, a data driven approach is used, harnessing millions of transaction records matched to questionnaires measuring demographics, shopping motivations (including questions on environmental concerns), and individual differences to investigate who is continuing to frequently purchase plastic bags after the levy.

The features were elicited from a survey of 12,986 respondents designed to study how psychological features relate to shopping behaviour. Survey responses were linked with till transaction records of actual plastic bag purchases. Using real world behaviour records doesn't allow us to control for confounding variables whilst the behaviour is being measured. Therefore, extraneous variables can be controlled for by inputting all available information into the model, to try to recreate the influences on behaviour in the real world. Deviating from the more traditional hypothesis driven approach allowed the exploration of a larger number of more complex explanations. Because of this exploratory approach particular caution is taken to ensure "p-hacking as procedural overfitting" (Yarkoni and Westfall, 2017)

does not occur (further discussed in the *Data Analysis* section).

6.2.2 Predictions

Despite testing all available variables, predictions were made on the expected predictors of frequent plastic bag purchasing based on two different components 1) the characteristics of people who intend to not consume plastic bags (Poortinga et al., 2016; Jakovcevic et al., 2014); and 2) the characteristics influencing people’s ability to convert intentions to behaviour (Hines et al., 1987; Bamberg and Möser, 2007).

Firstly, based on previous findings about the motivations for bag purchasing (Poortinga et al., 2016; Jakovcevic et al., 2014; Sun et al., 2017), it was expected that a person who was motivated to save money or had concern for the environment would intend to bring their own bags to the store instead of purchasing single-use ones. Therefore, it was predicted that frequent bag buyers were more likely to have a higher income, spend more money in store, be less frugal, and have less concern for the environment than infrequent plastic bag buyers.

Secondly, intending to reduce plastic bag consumption and actually doing so are not the same (Diamantopoulos et al., 2003; Bamberg and Möser, 2007). Many factors can disrupt a motivation or intention converting to behaviour (Kaiser et al., 1999; Sun et al., 2017). Bringing a personal bag to the store is often a required antecedent to refusing a plastic bag at the till, and requires forethought and planning. In other words, if someone was motivated to refuse plastic bags to save money or to help the environment, they could still end up buying bags because they had not adequately planned ahead or had purchased more items than intended on the spur of the moment. Thus, it was predicted that impulsiveness (measured using Behavioural Activation System (BAS) scales, and self-control) would positively predict regular plastic bag purchasing.

These hypotheses do not comprise exhaustively the factors being investigated. This research constitutes a bottom-up exploratory investigation into the demographic and psychological variables affecting single use plastic bag consumption. Consequently, this leaves the following sections open to report and discuss all findings (not just those attached to a particular hypotheses). Therefore, the available data is being exploited to explore the open ended question: who frequently buys plastic bags?

Two analyses were performed using two different methods of engineering the dependent variable. The first analysis uses the number of times a person bought a plastic bag to categorise regular bag buyers from non-regular buyers. This extracts a smaller sample of people from the transaction data. So the model is not just using the frequency at which people shop at Boots in the model, customers’ shopping frequency and amount spent is controlled for by inputting these variables into the model. The second analysis controls for shopping frequency and the quantity purchased at each transaction by engineering it into the DV. Instead of using just frequency, the DV measured the % of times a person bought a bag when their transaction data suggested they required one. The % cut-off chosen to determine

a frequent bag buyer created a slightly larger sample size. The two DVs had slightly variant predictors; the significance of this and the importance of engineering variables in big data is discussed.

6.3 Analysis 1

6.3.1 Method

Participants

Participants were from the sample of customers who's completed survey responses were matched to their loyalty card data, as described in Chapter 3, Section 3.2. To attain the final sub-sample used in this analysis, I first identified regular bag users, here defined as those who had bought a plastic bag at the retailer 4 or more times in the 9 months after the levy ($N=1440$). I then selected a random sample of customers who bought less than 4 plastic bags in this period ($N=1440$). This produced a balanced binary sample ($N=2880$) of more frequent and less frequent bag buyers.

Procedure

As outlined in Chapter 3, online survey included self-reported measures of personality, shopping motivations, and demographic questions. Before the survey began, participants provided permission - through informed consent - for researchers to access four years of their loyalty card transaction data, and entered their loyalty card number into the survey. No incentives were offered for taking part in the study. This study received full ethical approval from the university.

Transaction Data

Plastic bag sales were coded in the data for shoppers in England after the 5p levy was brought into action on the 5th October 2015. Although bags for life can also threaten our marine environments (DEFRA, 2018b), at this time they were seen as the pro environmental choice (Poortinga et al., 2016). Therefore, bags for life were not included in the analyses as the goal was to isolate those who were not impacted by the levy intervention. Therefore, the bags in this study are either single-use or the medium-strength plastic bags for which the levy applies. The total amount of transaction data that was linked to the survey respondents consisted of 1,284,825 distinct transactions from 1,222 different stores by 12,968 people. The total number of plastic bags bought by this sample was 20,504. From this data a smaller sub-sample was extracted, which had the same number of regular bag buyers and non-bag buyers ($N= 2880$) along with their survey responses, and their average monthly spending on their loyalty card. Therefore, for each of the 2,880 participants data was attained on personality, shopping motivations, and demographic variables (from the survey), and matched to data

from transactions on average monthly spending and whether they were a frequent bag buyer or not.

Materials

The survey used in this study was the generalised survey described in Chapter 3, which was designed with the intention of studying more generally which psychological factors affect different purchasing behaviours. In addition to the personality and demographic questions, a series of questions were asked about people’s shopping motivations, which included two questions on environmental concerns: ‘Environmental considerations affect the products that I purchase’; and ‘I am concerned about climate change (also known as global warming)’. It is worth noting that the questions and measures used in this study were not designed for the sole purpose of understanding plastic bag purchasing, and in this way could be considered secondary data. A link to a full copy of the survey sent to participants can be found in the Supplementary Materials (Section 10.1.1).

Data Analysis

Before being entered into the models, categorical variables were dummy coded, giving a total of 87 predictors at this stage. For income, the median household income group £50,000-£74,999 was the reference class. For the rest of the variables the most frequent class was the reference class: for loyalty card use this was ‘every time’, for shopping frequency ‘once a fortnight’, for the people you are living with this was ‘partner’, for marital status ‘married’, for occupation ‘intermediate manager/professional/administrative’, for highest qualification ‘bachelors’degree, and for region the most common was ‘Yorkshire’. To ensure the model was not simply capturing the characteristics of people who were buying more at the retailer, and thus requiring more bags, average monthly spend (when using their loyalty card) and self-reported shopping frequency (unrelated to loyalty card use) were included as controls. To aid interpretation, the continuous variables were mean centred and standardised before being input into the model.

The analysis utilised an inductive data-driven approach, a methodology common in computer science or other fields where mass datasets are available. This involved a typical machine learning paradigm testing different prediction models to see which model performed the best, and then extracting from the best model what the important predictors were. This method doesn’t require prior assumptions about the structure of the data or theoretical reasoning about possible variable relationships. Instead, it allows the data to tell us what variables are important predictors of plastic bag consumption. For a further discussion on the advantages of this approach compared to a traditional deductive data modelling methodology see Breiman et al. (2001).

The goal was to identify the variables in the data which were most predictive of regular plastic bag buying. To ensure useful information about the predictor variables was being gained, it was ensured the best, most accurate class of model available was used. Two classes

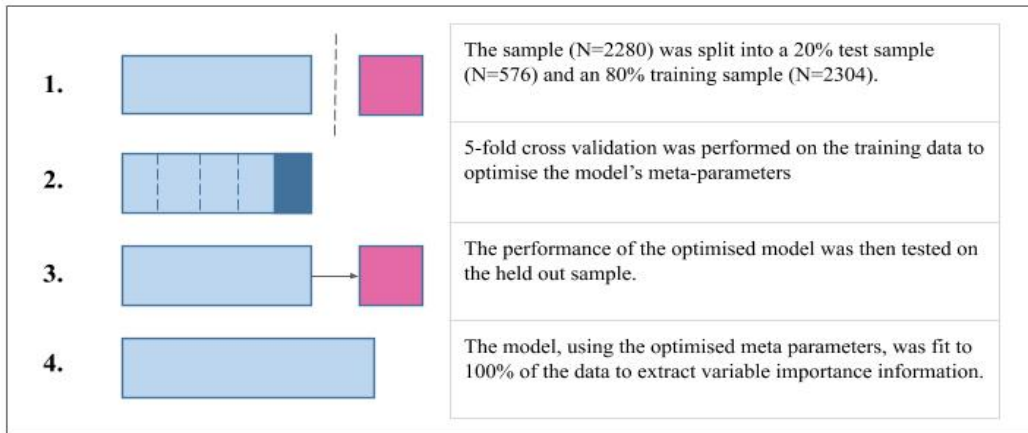


Figure 6.1: The cross-validation paradigm used in this study

of models were compared: a non-linear random forest (RF) model (Breiman, 2001) and a linear logistic regression model to see which gave the most accurate predictions.

Notably the decision to include a large number of variables in the model's selection pool brings with it an increased risk of over-fitting. Without care, a sample fit can be produced (i.e. β values/ variable importance observations) that is unrepresentative of the general population. This issue is known as 'procedural overfitting', (Yarkoni and Westfall, 2017) and an inadvertent form of p -hacking. Such dangers can be mitigated against through the rigorous use of *cross-validation* and, for Logistic regression, the use of regularisation parameters (Ng, 2004), and/or recursive feature elimination (Pedregosa et al., 2011). The key to the cross-validation approach (well established in computational science and of growing use in the traditionally inferential statistical sciences (Dahl et al., 2008; Yarkoni and Westfall, 2017)) lies in the iterative use of hold-out data partitions to ensure that the selected model (or more accurately model specification: in this case the independent features/model meta-parameters controlling regularisation) is the one that optimally generalises to out-of-sample data. A diagram illustrating the methodology and the datasets used for training and testing can be viewed in Figure 6.1.

First, 20% of the sample was excluded at random and set aside as a test set to evaluate the models on later. The remaining 80% of the data were used as training data. On this set the models would learn the patterns in the data (Bishop, 2006). To optimise the pattern recognition, different meta(hyper)-parameters can be tuned. The optimal parameters for each model were found using a grid search. The grid search algorithm repetitively evaluated a range of different combinations of possible parameter values, which were fed to the algorithm in a grid format. After trying all possible combinations, the algorithm selects the parameters which produce the most accurate model (evaluated here by F1 score). The training and evaluation of the parameters were done using K-fold cross validation on the training data (80% sample), where K=5 (Bishop, 2006). K-fold cross validation is used to recycle data for training and testing. The data is divided into K equal samples (folds). One fold becomes the test set, and the remaining folds are the training data. This process is repeated until each fold has been the test data. The mean error (or other accuracy statistic) from the K folds is taken as the performance of that particular set of parameters. The exact meta-

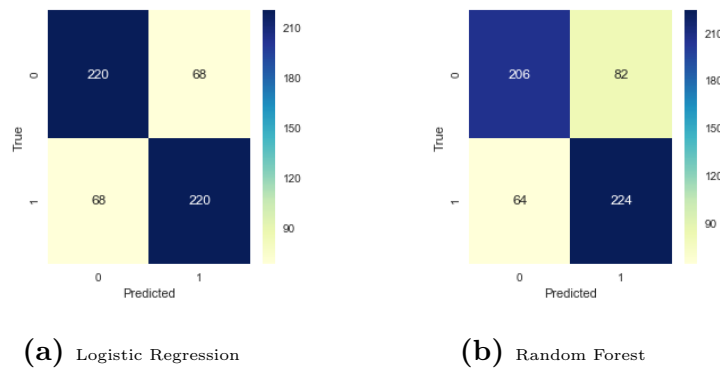


Figure 6.2: Confusion matrices for a) the logistic regression and b) the random forest models on the test data. 0 = non bag buyers, 1 = regular bag buyers.

parameters that were tuned for the logistic regression and the random forest can be found in the Supplementary Materials (Section 10.3.1).

Once both the random forest and logistic regression models had been optimised on the training data, their performance was then tested on the unseen (20%) test data. The models' accuracy was evaluated using the F_1 score. Testing the models on data that is unseen by the algorithm ensures that the model generalises and isn't over-fitting to the sample. The random forest had an F_1 score of 0.75, and the logistic regression's F_1 score was 0.76. The confusion matrices for the logistic regression and the random forest can be found in Figure 6.2. The confusion matrices show that the logistic regression was equally as good at predicting non bag buyers and regular bag buyers. The RF, however, was better at predicting the regular bag buyers than non bag buyers. Therefore less regular bag buyers were 'missed' by the RF, but there were more false positive errors. Due to the slightly superior accuracy of the logistic regression, the more balanced errors, and the principle of Occam's Razor (the RF is much more complex) the logistic regression model was chosen to extract the predictors of regular plastic bag buying.

To identify which variables were the important predictors in the logistic regression model standardised β coefficients and p -values were used. Before extracting the β coefficients, the logistic regression model was first fitted to 100% of the data to give the algorithm all the information available for the variable importance analysis. To prevent the model from over fitting at this stage, the same meta-parameters which were selected from the grid search were used (as these had produced a model which generalised well).

Multicollinearity was then checked for using the Variable Inflation Factor (VIF). Four variables had a VIF above the chosen threshold of 2.5 (Allison, 2012): 'Happiness' and 'Satisfaction with life' were highly correlated, as were 'Living with children', and 'Living with a partner and children'. To ensure the multicollinearity did not interfere with the variable importance measure, two variables were removed: 'Living with children' and 'Satisfaction with life' and refitted the model to 100% of the data, after which VIF levels were subsequently below the 2.5 threshold ('Satisfaction with life' = 1.44, 'Children' = 1.42). This threshold in the context of the large dataset is particularly stringent (O'brien, 2007). Furthermore, a Chi-square test showed that the new model with the two variables removed did not perform

differently from the old model (deviance = -0.04, $p=0.98$). A recursive feature elimination (Pedregosa et al., 2011) was also performed on the logistic regression inputs using cross-validation which identified one variable which could be removed without affecting the AUC score, indicating its redundancy. The variable dropped was ‘Duration’ measuring the time it took the person to complete the online questionnaire. This left a total of 84 predictor variables which were entered into the regression model.

6.3.2 Results

Descriptive Statistics

The descriptive statistics for the continuous demographic and personality variables can be found in Table 6.1. The plastic bag sales in England per 1,000 items from the whole survey sample ($N=12,968$) plotted by week over the 9 months after the levy can be viewed in Figure 6.4. Across the UK, more bags per 1,000 items were bought during the Christmas festive period (week 50-53: 7th December to the 3rd January). The 9 months following the levy start date (15th October 2015) plastic bags bought per week per 1,000 items in England show a significant linear decrease ($t(40) = -3.55, p = 0.001, R^2 = 0.24$). However, as have the data on plastic bag consumption prior to the levy, or a full years data after the levy, is not available it can not be known whether this decrease in bags is related to the levy, or an annual seasonal trend (more bags bought around Christmas and fewer in the summer).

Comparing the distribution of plastic bag sales across different regions within the UK (see Figure 6.5), Northern Ireland had the highest % of bags per 1,000 items bought, followed by Scotland, Wales, and regions of North West England. Regions in the South West of England (Cornwall and Devon) had bought the lowest number of bags per items. An ANOVA showed that there were significant differences in the number of plastic bags bought per 1000 items in different regions shown in Figure 6.5 ($F(11, 1210) = 8.8, p<0.001, \eta_p^2 = 0.07$).

Bonferroni corrected t -tests showed that in England significantly fewer plastic bags were bought per 1,000 items than in Northern Ireland ($t(36)=-3.23, p<0.001, d=0.85$), Scotland ($t(115)=-4.03, p<0.001, d=0.65$) and Wales ($t(62)=-3.00, p<0.001, d=0.87$) but there were no statistically significant differences between the latter three countries. A multi-level model with random intercepts for store type showed that the country effect remained when the type of store (i.e. chemist, high street (small or large), convenience, or travel store) was accounted for ($F(3,1211) = 26.40, p<0.001$).

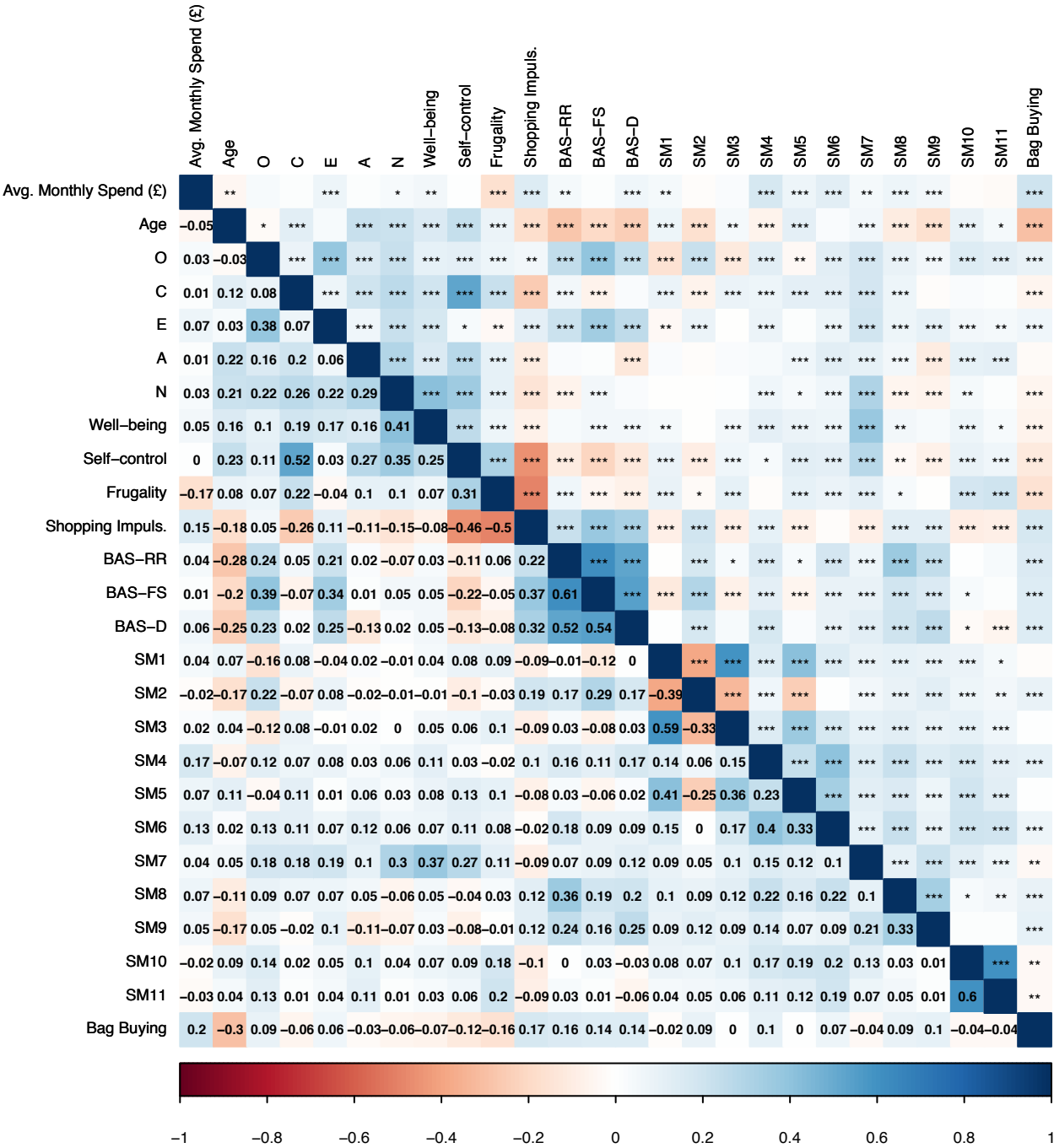


Figure 6.3: Correlation matrix of the continuous independent variables and the dependent variable ‘Bag Buying’ where 1 = frequent bag purchasing and 0 = infrequent. The lower section of the matrix displays the Pearson r correlation coefficients. The upper section shows the significance levels where *p<0.05; **p<0.01; ***p<0.001. SM1: ‘I would rather stick to a brand I usually buy...’, SM2: ‘I enjoy taking chances in buying unfamiliar brands...’, SM3: ‘If I like a brand I rarely switch...’, SM4: ‘I would not mind paying more for high quality...’, SM5: ‘I only buy products I trust.’, SM6: ‘Product quality is extremely important to me.’, SM7: ‘I am pleased about the way I look...’, ‘SM8: When I feel good about my looks, I am happier...’, SM9: ‘When I see a mirror, I ”check myself out”...’, SM10: ‘Environmental considerations affect the products that I purchase.’, SM11: ‘I am concerned about climate change...’

Prediction Results

To investigate who regularly bought plastic bags, a logistic regression model was optimised on training data (see Methods) to predict regular bag buyers versus non-bag buyers in an

unseen test dataset. The independent variables were self-reported demographics, personality traits, and shopping frequency/motivation information from the questionnaire, as well as participants' average monthly spending on their loyalty cards. The logistic regression, when tested on the 20% balanced unseen data ($N=576$), classified regular bag buyers and non bag buyers correctly in 76% of cases (F_1 score = 0.76) (precision values were 0.76 for both classes).

When the model was fitted to the whole sample of data to extract variable importance, the McFadden R^2 value (McFadden et al., 1973) for the model was 0.29. All 15 of the significant predictors (along with Reward Responsiveness (BAS-RR) which was borderline), their direct correlations with the dependent variable, standardised β regression coefficients, and p -values can all be viewed in Table 2. This left 68 variables that were not significant. These included climate change concerns, gender, four of the Big Five personality traits (C, E, A, N), and self-reported geographic region. A full list of these variables can be found in the Supplementary Materials (Section 10.3.2).

Variable Importance Results

The standardised β coefficients and p -values from the logistic regression (fit to 100% of the sample) were used to measure the importance of the independent variables in the model predicting regular/non bag buyers. Shopping frequency, amount spent, impulsivity, a present time focus (preferring £55 now to £75 in 3 months), and Openness (O) positively significantly predicted regular bag buying. Whilst age, frugality, and the income groups £25,000-£34,999 and £35,000-£49,999 negatively predicted bag buying. Consumer environmental considerations significantly negatively predicted bag buying, but climate change concern did not. BAS-RR significantly correlated with regular bag buying but was borderline significant in the logistic regression model ($p=0.054$).

Income

The number of regular bag buyers and non bag buyers for each of the income groups can be found in Figure 6.6. There are more non-bag buyers in the lower income groups (less than £49,999) and 'prefer not to say'. After £50,000 there are more regular bag buyers than non bag buyers. Therefore, around the amount of £50,000 annual household income appears to be the switch over point from more non bag buyers to more regular bag buyers.

In the regression model the group £35,000 to £49,000 is a small but significant negative predictor of regular bag buying (when compared to the median reference class of £50,000-£74,999), however Figure 6.6 shows that there is a similar number of regular ($N= 259$) and non bag buyers ($N= 261$) in this category. This suggests that this income category is predictive once the other variables in the model have been used to categorise regular and non bag buyers.

Mediation

With respect to the direct correlations in Table 6.2, Fun Seeking (BAS-FS) positively correlated with bag buying, but in the logistic regression negatively predicted it. Furthermore, contrary to hypotheses Drive (BAS-D) was not significant in the regression and BAS-RR only borderline significant. Therefore, it was speculated that age could be masking some of the effect of impulsivity on bag buying. It was also theorised that income and frugality could also be related to age, partly given differences observed between direct correlations and β coefficients in the logistic regression. It was therefore hypothesised that as well as age directly predicting bag buying, frugality, impulsivity, and income would mediate the relationship between age and bag buying.

To test this theory Structural Equation Modelling (SEM) was used to analyse three indirect mediation pathways. For this analysis, income group was observed directly from the questionnaire, and coded as a categorical variable. Impulsivity was a latent variable derived from the mean of BAS-D, BAS-FS, and BAS-RR items on the questionnaire. Frugality was also a latent variable derived from the mean of 8 questionnaire items (see the individual items in the full questionnaire provided in the Supplementary Materials, Section 10.1.1). Covariances were included in the model as it was hypothesised that income and frugality would, to some degree, be related.

The pathways were tested for significance using 10,000 bootstrapping procedures. A Comparative Fit Index (CFI) of 1 and Root Mean Squared Error (RMSE) of 0.00 indicated a good model fit. The path diagram can be viewed in Figure 6.7. The direct effect of age on plastic bag buying, and all three indirect paths, were statistically significant. There were also significant covariances between the mediator variables impulsivity and frugality, and income and frugality, showing that the mediator variables are not entirely independent. A full table of the regression, covariance and pathway statistics can be found in the Supplementary Materials (Section 10.3.3).

Table 6.1: Descriptive statistics for the relevant continuous variables.

	Min	Max	Mean	SD
Age	17	93	46.68	14.99
Openness	1	7	4.91	1.16
Conscientiousness	1	7	5.76	1.11
Extraversion	1	7	4.04	1.56
Agreeableness	1	7	5.47	1.13
Emotional Stability	1	7	4.62	1.48
Happiness	0	10	7.11	2.10
Frugality	1	6	4.48	0.90
BAS-RR	1	4	3.25	0.47
BAS-FS	1	4	2.80	0.52
BAS-D	1	4	2.49	0.64
Environmental considerations...	1	7	4.41	1.50
I am concerned about climate...	1	7	5.01	1.57
Average monthly spend (£)	2.29	939.17	53.26	37.87

Note: The environmental shopping motivation questions were ‘Environmental considerations affect the products I purchase’, and ‘I am concerned about climate change (otherwise known as global warming)’ where a higher value indicated a greater level of agreement for each statement.

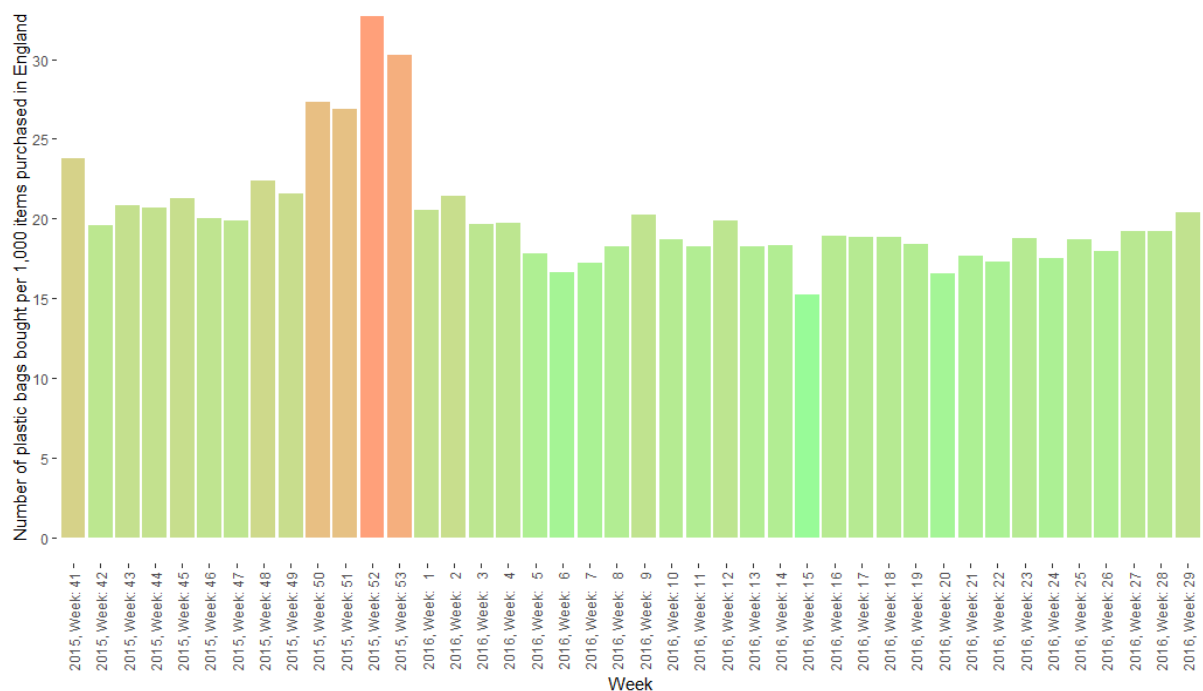


Figure 6.4: A graph to show the number of sales of single use plastic bags per 1000 items purchased in England each week by the survey sample ($N=12,968$). Week 44 is the week of Halloween, week 48 the week black Friday occurred, and week 52 is the week of Christmas.

6.3.3 Discussion

Descriptives

Overall a very small sample of people were frequently buying plastic bags, with over half of the full survey sample having not bought a single-use plastic bag in the 9 months after the levy. This is very encouraging, corroborating evidence from DEFRA (2018b) that the levy is working to reduce single use plastic bag consumption in the UK. Statistics on plastic bag use concentrate on bags bought from the top seven supermarket chains (DEFRA, 2018b) and less is known about the bag sales from non-supermarket retailers. The findings suggest that the success of the levy in reducing single-use plastic bags extended to high-street health and beauty retailers as well as supermarkets.

In the 9 months following the levy, the sales of plastic bags per items bought significantly decreased. However, because the data did not extend to after the 9 months it could not be concluded whether the trend was levy related or a seasonal bag purchasing trend (more around Christmas, and less in the summer). Consumers may have been buying more plastic bags around Christmas for practical reasons, for example, they didn't want others to see their gift purchases, or it might have made returns easier. Alternatively, the increase around Christmas could have reflected a change in customers' psychological states/processes. For example, greater present buying may have lead to a greater focus on materialism, which could have driven high consumption high throw-away behaviour. This is supported by average monthly spend being the strongest predictor (after age) of regular plastic bag consumption in the logistic regression model. Alternatively, people may have been more stressed over the festive period, leading to greater disorganisation, and less planning of shopping trips.

Table 6.2: Logistic regression model predicting regular plastic bag consumer in Analysis 1

	<i>Regular plastic bag buyers</i>		
	Pearson r	β coefficients	Odds Ratio
1. Average Monthly Spend (£)	0.33***	1.84*** [1.57, 2.11]	2.51
2. Age	-0.37***	-1.81*** [-2.08, -1.53]	0.41
3. Shopping Frequency: Once every 4-6 months	-0.09***	-0.80*** [-1.18, -0.42]	0.01
4. Shopping Frequency: Once a month	-0.22***	-0.77*** [-0.97, -0.57]	0.39
5. Shopping Frequency: Once every 2-3 months	-0.16***	-0.67*** [-0.90, -0.44)	0.21
6. Frugality	-0.20***	-0.46*** [-0.69, -0.23]	0.79
7. Income: £25,000-£34,999	-0.06**	-0.35** [-0.59, -0.10]	0.63
8. BAS-FS	0.09***	-0.33* [-0.61, -0.05]	0.85
9. Shopping Frequency: Once a week	0.16***	0.32** [0.12, 0.52]	1.44
10. £55 now, rather than £75 in 3 months	0.07***	0.30* [0.05, 0.55]	1.35
11. Openness	0.05* ($p=0.011$)	0.29** [0.07, 0.51]	1.16
12. ‘Environmental considerations...’	-0.06***	-0.27* [-0.50, -0.03]	0.87
13. Occupation: Homemaker	0.04 ($p=0.054$)	0.26* [0.06, 0.46]	3.43
14. BAS-RR	0.15***	0.25 ($p=0.054$) [-0.01, 0.51]	1.13
15. Income £35,000-£49,999	-0.002	-0.24* [-0.48, -0.01]	0.73
16. Qualification: Postgraduate Diploma	0.03	0.23* [0.02, 0.43]	1.42
Observations			2,880
Log Likelihood			-1,415.90
Akaike Inf. Crit.			2,999.80

*Note: Significance levels are * $p<0.05$; ** $p<0.01$; *** $p<0.001$. ‘Environmental considerations...’ = ‘Environmental considerations affect the products I purchase’. The coefficients reported in the table are the standardised β coefficients with standardised 95% confidence intervals. The significant variables are ranked by the size of the absolute β coefficients. The β coefficient represents the degree of change in the dependant variable for every 1-unit (standard deviation) of change in the predictor variable.*

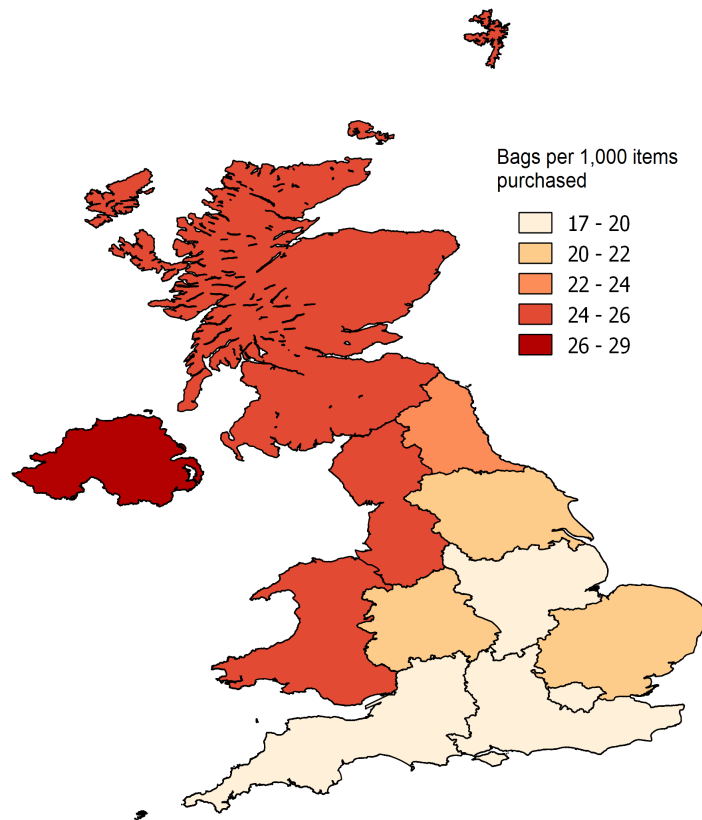


Figure 6.5: A map of the UK showing in each region the standardised percentage of bags sold by the retailer per number of transactions in that region. The standardised bag sales were calculated using the full survey sample of 12,968 people.

Overall, these findings suggest that plastic bag reduction campaigns targeted during the festive period could have considerably more impact.

The geographic distribution of the bag sales in the full sample, per number of items bought, showed evidence of a North-South divide in the UK. People in Northern Ireland, Scotland, the North West, and Wales purchase more bags than people in Southern regions. At a country level, England had significantly fewer bag sales per one thousand items than Northern Ireland, Scotland, and Wales. This shows the short-term effectiveness of the levy as an intervention in England, but highlights that an indifference to the levy in the other countries, which introduced the levy earlier, could have occurred. Wales introduced the 5p levy in 2011, then Northern Ireland and Scotland followed in 2013 and (October) 2014 respectively. National statistics showed a large decrease in bag sales in these countries initially post-levy (DAERA, 2018; Scottish Government, 2015; Welsh Government, 2016), but after a number of years the levels are significantly higher than that in England recently post-levy.

Some variants exist in the levies across the different UK countries in terms of price, and also the type of bag the charges applied for. England, for example, was the only country where only plastic bags were charged, whilst for the other countries the levy applied to paper bags as well (Thomas et al., 2019). Still, it is possible the results indicate that the short term effects of the levy are much greater than the long term effects. In (southern) Ireland a plastic bag levy was introduced in 2002, and initially bag use reduced by 94% (Convery et al., 2007). However, after the initial drop plastic bag use started to rise again (Clarke,

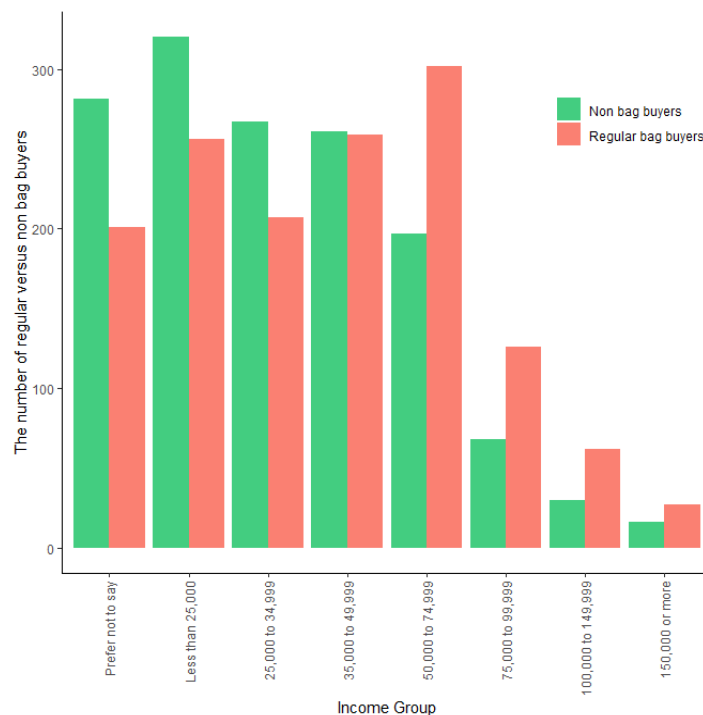


Figure 6.6: Number of bag buyers and non bags buyers in each income category for the balanced sample used in the logistic regression analysis ($N=2,880$)

2014; Xanthos and Walker, 2017). The data may be showing some indication of an increased indifference to bag charges in Wales, Northern Ireland, and even Scotland just a year later.

Prediction Results

It was hypothesised that frugality, income, and environmental shopping motivation questions would negatively predict bag buying behaviour - reflecting economic and ecological motivations. In the logistic regression model, 15 variables significantly contributed to the prediction of regular plastic bag buying. Frugality and lower income group membership were indeed found to be strong negative predictors of bag buying, supporting the hypothesis that people wouldn't buy plastic bags because they were conscious of money. The relationship between frugality and bag purchasing suggests that placing more value on thriftiness in society could improve adherence to plastic bag reduction strategies. The income group £25,000 to £34,999 negatively predicted frequent bag buying. This stands against literature suggesting that pro-environmentalists tend to be middle or upper-middle class individuals (Balderjahn, 1988; Howard et al., 1993). Despite Figure 6.6 showing that the lowest income group (less than £25,000 a year) had the most non bag buyers, it was not a significant negative predictor of bag buying in the regression model. This may indicate that other predictive factors, such as age, dominate here.

To evaluate the ecological motivations for bag buying behaviour, two questions were used: one on consumer environmental considerations and the other measuring climate change concern. Contrary to hypotheses, climate change concern was not a significant negative predictor of bag buying, however, consumer environmental consideration was. This demonstrates that there could be a disconnect between people's concerns on climate change and the act of buying a plastic bag. Plastic consumption until recently has been viewed as a separate issue to

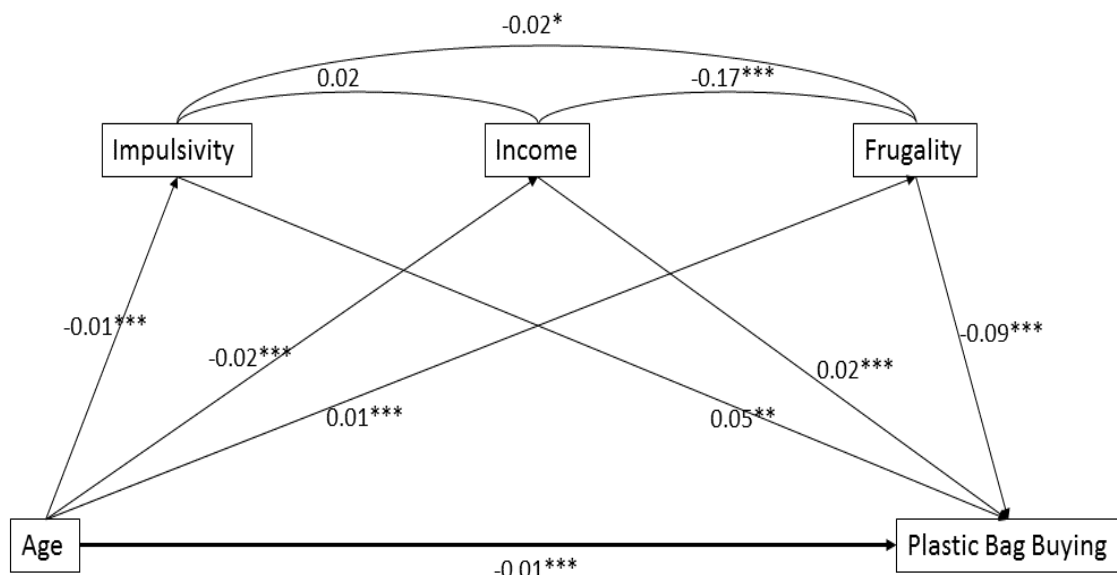


Figure 6.7: Standardised β coefficients from the SEM showing a direct effect of age on bag buying, three different indirect mediation pathways (through impulsivity, income and frugality), and the covariances between the mediators. For the full regression, covariance, and pathway statistics see the Supplementary Materials (Section 10.3.3)

greenhouse gasses causing global temperatures to rise (Royer et al., 2018). Recent research has now connected the two environmental issues revealing that as light breaks down plastic, two of the most problematic greenhouse gases, methane and ethylene, are released (Royer et al., 2018). More education on the relationship between plastic bag consumption and climate change could help reduce the number of plastic bags being bought - as those who are concerned about climate change may be more motivated to save buying plastic bags.

The final prediction was that the trait impulsivity (measured using BAS scales, or via a lack of self-control) would positively predict regular plastic bag buyers. It was reasoned that the propensity to plan ahead (synonymous with low impulsivity) is important for bringing personal bags to the store, and thus refusing plastic bags at the till. The findings showed that reward responsiveness (BAS-RR) positively related to regular bag purchasing, but the effect was borderline significant. The relationship between BAS-RR and bag consumption fits with the hypothesis: people who are highly responsive to rewards are more likely to make purchases spontaneously (or buy more than anticipated), and thus are less likely to have personal bags with them at the store. Drive (BAS-D) was also not a significant predictor; and contrary to expectations fun-seeking (BAS-FS) significantly *negatively* predicted bag purchasing. It was speculated that these diverging relationships between the BAS facets may be due to BAS-FS being less a material and more experience focused facet of impulsivity. For example, the fun seeking item ‘I crave excitement and new sensations’ suggests a greater need for experiential rather than materialistic purchases. Therefore, someone who is likely to spend their time and money seeking experiences rather than possessions will have less need for plastic bags.

In addition to impulsivity, it was speculated that other variables tapped into a propensity to plan ahead, such as those measuring forward/future thinking. For instance, people

answering affirmative to the time preference question ‘I would rather have £55 now than £75 later’ (positively related to regular bag purchases) are exhibiting a greater preference for present over future outcomes. Furthermore, some of the items on the frugality scale (negatively related to bag purchases) also reflect a future time preference. For example, the question ‘there are things I resist buying today so I can save for tomorrow’ is tapping into immediate impulse control for a greater reward in the future. People who orient their thoughts to the future are more likely to plan their shopping, increasing the likelihood that they will bring personal bags to the store. Furthermore, holding the future in greater esteem than the present may lead future focused people to consider the effect of today’s actions on tomorrow’s environment, in line with theory of psychological distance (Spence et al., 2012). Taken together, the results imply that it is not just (a lack of) impulsiveness that is important for predicting bag free shopping - it is having a cognitive orientation towards future thinking.

Despite the literature finding that the traits O, C, A, and N significantly predict environmental concern and engagement (Milfont and Sibley, 2012; Hirsh, 2010), it was found that only O significantly predicted plastic bag buying. The results showed that people who are open to new experiences are more likely to buy plastic bags - the opposite to what was expected from the literature showing high O individuals are more concerned and engaged with the environment (Milfont and Sibley, 2012; Hirsh, 2010). This could be either because there is something uniquely different about plastic bag buying and its drivers, compared to other ecological activities; or because real world behaviour was measured, and this is somewhat removed from self-reported behaviours and environmental concern/engagement (Kormos and Gifford, 2014). As environmental concerns themselves were not predictive of the bag buying behaviour, the latter is imagined to be the more likely explanation.

Age was a strong predictor of bag buying where older customers, were less likely to be regular bag buyers - replicating Poortinga et al. (2016)’s findings. Although trait theory suggests that personality remains stable over time (Cobb-Clark and Schurer, 2012), the structural equation model (SEM) showed that age negatively related to impulsivity (measured using BAS) and positively to frugality. Thus, it is suspected that people who are older are better at planning ahead and bringing personal bags to the shops. These findings corroborate previous research suggesting that future plastic bag reduction efforts should concentrate on targeting younger people, with interventions or plastic bag alternatives to make it easier to make a plastic free choice at the till.

The SEM also revealed that impulsivity, frugality, and income mediate the relationship between age and bag consumption. This illustrates that it is not only age itself that affects bag buying but its co-occurring or correlated psychological traits (impulsivity and frugality) and income level. The model additionally highlighted the inter-connectivity of the variables, as the mediators were not just significantly related to age and bag consumption, but also to each other (with the exception of income and impulsivity). The SEM identified the significant relationship between impulsivity and bag consumption, without the direct effect of age -

which was speculated to be masking BAS's effect in the regression model. This highlights one of the difficulties of data with many features, and why follow up analyses to aid the interpretation of prediction models, can be an important part of the process of extracting useful information from big data. It is, however, recognised that the SEM is a simplification of a complex network of interrelated variables and pathways - and certainly not the whole picture.

6.4 Analysis 2

In Analysis 1, spending and shopping frequency (along with age) were the greatest predictors of regular plastic bag buying. This could either be because people who shop frequently are more materialistic/consumerist, and are therefore more likely to be involved in a throw away culture of disposable bag buying; or people who shop more frequently are purchasing plastic bags for a similar proportion of shopping trips than non-frequent shoppers (just their total number of bags purchased is greater as they visit the store more often). To investigate, the present analysis I engineered a new dependent variable to tease apart these different drivers of plastic bag consumption, and measure the bag purchasing as a factor of the number of shopping trips.

Fifteen variables out of eighty-four were found to be significant in Analysis 1. Conducting eighty-four separate significance tests makes it more likely that some of these variables are going to be significant by chance. Therefore, some of these variables, particularly those with wide confidence intervals, may be false positive results. This raises the issue of how to encourage knowledge discovery using an inductive approach, whilst ensuring nothing resembling a "fishing" exercise (where researchers continue to 'recast their net' until they find something significant) is being carried out? In this analysis, a more stringent criteria of significance is applied to control for a False Discover Rate (FDR) of 5%.

To summarise, Analysis 2 is conducted the same as Analysis 1, but with two methodological differences. First, the dependent variable is engineered to investigate a conceptually slightly different question of who buys plastic bags 25% or more of the trips when they have bought enough items/volume of items to warrant needing one? This refinement reduces the effect of shopping frequency on the outcome variable, putting us in a position to better answer the question - what are the psychological and demographic drivers of single-use plastic bag consumption? Secondly, the FDR is controlled for when assessing the significance of important variables. If psychological or demographic variables are consistently significant across both Analysis 1 and Analysis 2, and in spite of the more stringent significance threshold applied in Analysis 2, stronger conclusions can be drawn on the robustness of these features as predictors of the real world ecological behaviour, plastic bag buying.

6.4.1 Method

Participants

The same data pool of 12,968 survey responses matched to loyalty card data that was used in Analysis 1, was used for Analysis 2. However, the definition of frequent bag buyers was slightly different (see Section 6.4.1), leading to a variant subsample of frequent and non-frequent bag buyers being selected. A balanced binary sample of 4,652 was used for this analysis.

Procedure

The total quantity of transaction data which was linked to the survey participants (with complete questionnaires) consisted of 1,284,825 distinct transactions from 1,222 different stores by 12,968 people. Using this data, a percentage was calculated of the times each person bought a plastic bag when the data suggested they had bought a large enough quantity of items to warrant needing one. This was done by only analysing the shopping trips where either one the following conditions was satisfied: 1) they had 3 or more items in their ‘basket’ (used here to denote the items bought in one transaction, not a physical basket), or 2) they had a basket volume of over 2,225 ml (approximately an 8th of the total volume of a standard carrier bag volume (18 litres)).

By only analysing these higher-quantity baskets the selection criteria does not include visits where a customer buys only a couple of items that could fit in their pocket/hands and thus dose not require a plastic bag. Therefore, the measure of bag buying is the proportion of times a person makes a single-use plastic bag purchase, when a bag was deemed to be required to carry their shopping. If a bag was not brought, when the conditions for needing one were satisfied, it was inferred they either bought their own bag to the store, made an effort to carry multiple items or a large volume of items by hand, or bought a bag for life.

Creating a Balanced Modelling Sub-sample

From the whole sample ($N=12,968$) 56% had never bought a single-use plastic bag. Of those who had bought a bag at least once, the proportion of purchases that included a single-use bag when they had actually bought enough items/volume of items to suffice needing one is plotted in Figure 6.8. Only 1,014 people had bought a bag 50% or more of their bag-warranting visits. Because of the non-normal distribution (positive-skew) of bag buying behaviour, rather than predict the percentage of trips a bag was bought using a linear regression, I chose to create a binary outcome variable which divided people into frequent and infrequent bag buyers by choosing a cut-off percentage value. When choosing this value I weighed up a) extracting a sample of people who are repeatedly buying plastic bags, with b) the size of the sample attained (i.e. avoiding throwing data away). It was decided that, because of the strong positive skew (see Figure 6.8), the cut-off of buying a bag 25% or more of the bag-warranting trips to the store was sufficient to separate frequent bag buyers from

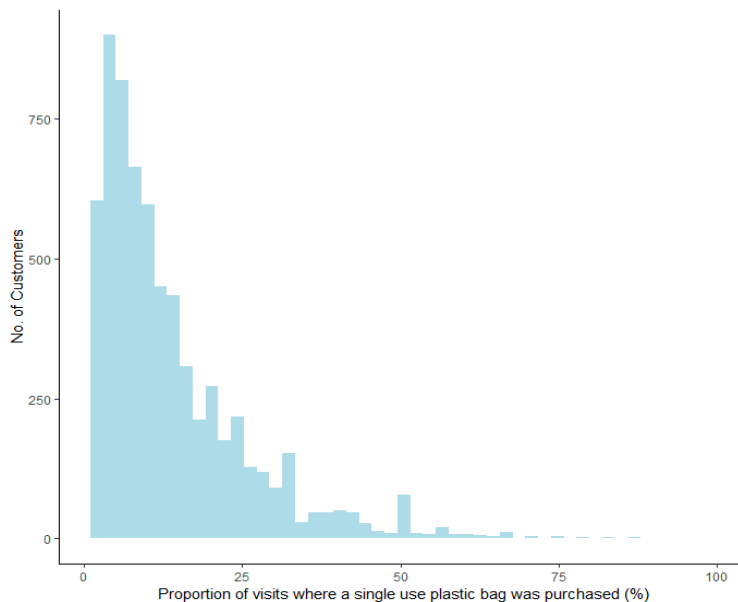


Figure 6.8: A graph to show the distribution of plastic bag purchases when enough items or volume of items to warrant needing one were purchased. Those who never purchased a bag (0%) ($N=8,546$) are not represented in this graph. The cut-off of 25% or more determined the frequent bag buyers.

less frequent bag buyers (the norm in the data).

Therefore, individuals who bought bags on 25% or more occasions, and who had completed questionnaires, were deemed the frequent bag buyers and made up the positive class in the modelling sub-sample ($N=2,326$). I then randomly sampled an equal number of people who bought a plastic bag with less than 25% of their bag warranting visits - producing a balanced sample of 4,652 for modelling. Therefore each individual in the final sample had an outcome label of either being a frequent bag buyer or not (the dependent variable). Participants' spending data was then used to calculate their average monthly spending, and used personality information, shopping motivations, and demographic variables collected from the survey as independent variables. The descriptives of the continuous independent variables from the balanced modelling sample are shown in table 6.3.

Data Analysis

This analysis considers only the logistic regression model to predict the binary outcome variable of more frequent versus less frequent bag buying. This decision was made for 2 reasons: Firstly, the logistic model had already shown to be at least as good as the random forest in Analysis 1; and secondly, I wanted to extract the *statistically* significant important predictors of frequent bag buying.

Before being entered into the model, all categorical variables were dummy coded, giving a total of 92 predictors. Contrary to Analysis 1, which used the median income group as the reference class, the lowest income category 'Less than £25,000' was removed. As before, for the rest of the variables the most frequent class was the reference class, in order to compare the effects of smaller groups to the larger normative group.

Although shopping frequency is being controlled for (by engineering the outcome variable

using the percentage of times a person bought a bag when one was required), it was further controlled for by including self-reported shopping frequency, loyalty card use, and average monthly spend (calculated from the transaction data) as features in the model. To aid interpretation, all continuous variables were mean centred and standardised before being input into the model. The same procedure of meta-parameter optimisation, using a grid search and 5-fold cross validation was used as in Analysis 1 and outlined in Figure 6.1 (see the Supplementary Materials, Section 10.4.1 for the meta parameters used).

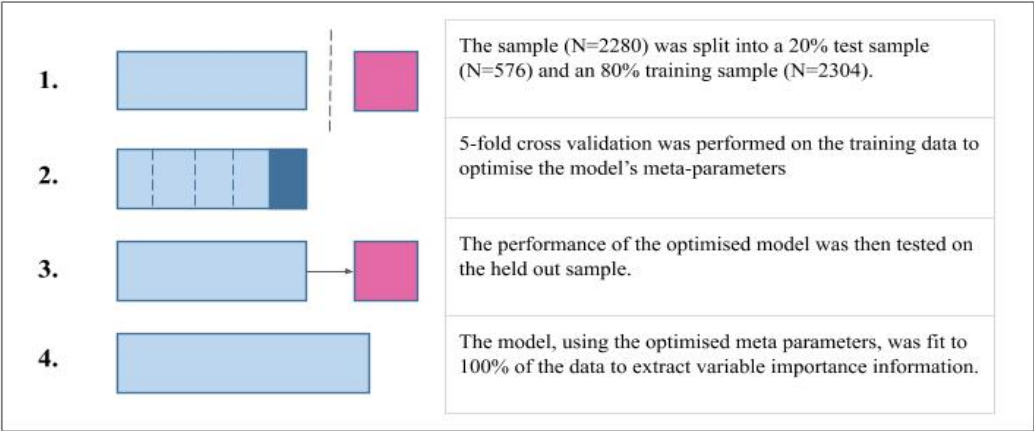


Figure 6.9: The cross-validation paradigm used in this study to prevent *p*-hacking

As a result of the model selection procedure above, I am left with a single regularised logistic regression model with fixed meta-parameters and input features, selected so as to avoid *p*-hacking as procedural overfitting. At this point I return to a more traditional analysis when investigating the significance of the predictive variables via standardised β values, controlling for the likelihood of some variables being significant by change using the Benjamin-Hochberg false discovery procedure¹.

Before interpreting the standardised β values in the context of variable predictive importance, as in Analysis 1, I first checked for multicollinearity between the variables using the variable inflation factor (VIF). Four variables had a VIF above the chosen threshold of 2.5 (Allison, 2012): ‘Happiness’ and ‘Satisfaction with life’ were highly correlated, as were ‘Living on own’, and ‘Marital Status: Single’. Rather than removing the variable ‘Satisfaction’, as in Analysis 1, it was combined with ‘Happiness’ into a single Well-being variable (Diener et al., 2002), whilst ‘living on own’ was removed. The model was re-fit to check the VIF levels and all were subsequently below the 2.5 threshold (‘Well-being’ = 1.45, ‘Marital Status: Single’ = 1.64) (O’Brien, 2007). As before, the final optimised mode was then tested on out-of-sample data using the F1 score.

The final step of the analysis was to extract the significant predictive variables of frequent bag buying behaviour from the model. Due to having a large number of independent variables, unlike in Analysis 1, it was additionally controlled for the likelihood of some variables being significant by chance by using the Benjamini–Hochberg false discovery procedure with

¹It is noted that the use of cross-validation can theoretically be extended to address issues of *p*-hacking within variable importance calculations. However, its use is not widespread and as such more traditional approaches are preferred in this work.

a q -value of 0.05 (Benjamini and Hochberg, 1995). The q -value is the expected proportion of false positives among all features as or more extreme than the observed one. Therefore a q -value threshold of 0.05 yields a FDR of 5% among all features called significant. This is not to be confused with a p -value threshold (alpha) of 0.05, which yields a false positive rate (FPR) of 5% among all *truly null* features. By only reporting variables which passed the Benjamini–Hochberg test, the likelihood of type 2 error (missing a positive result) was increased, in order to reduce the likelihood of a type 1 error (false discovery) (Benjamini and Hochberg, 1995).

6.4.2 Results

The logistic regression, when tested on the 20% balanced unseen data ($N=930$), classified frequent bag buyers and infrequent bag buyers correctly in 67% of cases (F_1 score = 0.67). The model was equally good at classifying frequent and infrequent bag buyers (see Figure 6.10).

When the model was fitted to the whole sample of data to extract variable importance, the McFadden R^2 value (McFadden et al., 1973) for the model was 0.13. 25 variables were significant predictors of bag buying frequency with a threshold of $p \leq 0.05$. After the Benjamini–Hochberg false discovery procedure ($q=0.05$, see *Methods* section) 10 significant predictors of frequent bag buying remained: Age (negatively), Average monthly spend (positively), Gender (Male positively), Frugality (negatively), Openness (positively), the shopping motivation question ‘I am pleased with the way I look’, and the self-reported “home” geographical regions ‘West Midlands’, ‘North West’, ‘North East’, and ‘East Anglia’.

The 10 predictors and their direct correlations with the dependent variable, standardised β regression coefficients with 95% confidence intervals, and p -values can all be viewed in Table 6.4. Notably the two questions measuring environmental concerns were not significant, and neither were impulsivity measures. A full output of all the logistic regression features, their β coefficients, and p -values can be found in the Supplementary Materials (Section 10.4.2).

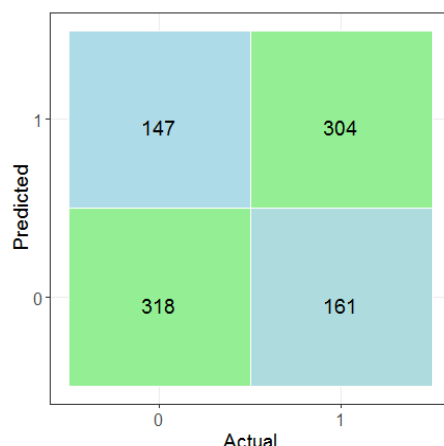


Figure 6.10: Confusion matrix for the logistic regression predicting the test data. 0 = infrequent bag buyers, 1 = frequent bag buyers.

Table 6.3: Descriptive statistics for the continuous independent variables

	Min	Max	Mean	SD
Average Monthly Spend (£)	2.67	380.44	42.40	27.35
Age (years)	17	95	50	15
Extraversion	1	7	4.05	1.55
Agreeableness	1	7	5.51	1.12
Openness	1	7	4.85	1.19
Conscientiousness	1	7	5.81	1.10
Emotional Stability	1	7	4.69	1.47
Well-being	0	10	7.18	1.89
Self-control	1.15	5	3.31	0.66
Frugality	1	6	4.56	0.88
Shopping Impulsivity	1.11	5.11	2.73	0.86
BAS-RR	1.20	4	3.22	0.46
BAS-FS	1	4	2.77	0.53
BAS-D	1	4	2.44	0.64
SM1: I would rather stick to a brand I usually buy...	1	7	4.51	1.57
SM2: I enjoy taking chances in buying unfamiliar brands...	1	7	3.97	1.50
SM3: If I like a brand I rarely switch...	1	7	4.65	1.54
SM4: I would not mind paying more for high quality...	1	7	5.43	1.23
SM5: I only buy products I trust.	1	7	5.26	1.40
SM6: Product quality is extremely important...	1	7	6.12	1.05
SM7: I am pleased about the way I look...	1	7	4.32	1.54
SM8: When I feel good about my looks, I am happier...	1	7	5.78	1.19
SM9: When I see a mirror, I "check myself out"...	1	7	3.90	1.75
SM10: Environmental considerations affect the products that I purchase.	1	7	4.43	1.52
SM11: I am concerned about climate change...	1	7	5.05	1.55

Note: SD = Standard Deviation. SM refers to Shopping Motivation, so that SM1 refers to the first item on the Shopping motivation questionnaire (and so on). For the full SM items see the full regression table in the Supplementary Materials (Section 10.4.2).

Table 6.4: Logistic regression model predicting frequent plastic bag buyers in Analysis 2

	<i>Frequent plastic bag buyers</i>		
	pearson <i>r</i>	β coefficients and 95% CIs	Odds Ratio
Age (years)	-0.30	-0.60***[-0.70,-0.50]	0.96
Average monthly spend (£)	0.20	0.43***[0.35,0.51]	1.02
Gender: Male		0.21***[0.15,0.28]	2.15
Frugality	-0.16	-0.21***[-0.29,-0.13]	0.78
Openness	0.09	0.14***[0.06,0.22]	1.12
Region: West Midlands		0.12**[0.05,0.19]	1.68
SM7: I am pleased about the way I look	-0.04	-0.13**[-0.20,-0.05]	0.92
Region: North West		0.12**[0.04,0.19]	1.51
Region: North East		0.11**[0.04,0.18]	1.60
Region: East Anglia		0.10**[0.03,0.17]	1.68

*Note: Significance levels are * $p<0.05$; ** $p<0.01$; *** $p<0.001$. The coefficients reported in the table are the standardised β coefficients with standardised 95% confidence intervals. The pearson r coefficient is provided to show the correlation with the outcome variable for continuous independent variables only. The significant variables are ranked by the size of the absolute β coefficients. The β coefficient represents the degree of change in the dependant variable for every 1-unit (standard deviation) of change in the predictor variable.*

6.4.3 Discussion

This second analysis was conducted to address two methodological issues in Analysis 1: that 1) the model may be predicting shopping frequency, rather than the frequency of bag buying; and 2) that some variables may be significant by chance. The re-engineered dependent variable prevented shopping frequency from being included in the measure of frequent bag buying. Although this reduced the prediction accuracy of the model, it allowed for a clearer picture of the psychological and demographic predictors of plastic bag consumption - rather than consuming more in general. The prediction model did not find any of the shopping frequency variables significant. Average monthly spend remained an important predictor, but was no longer the best predictor - age was. This suggests that shopping frequency is not a predictor of plastic bag buying, yet spending more is. The latter is likely to be a psychological disposition for spending more (linked to frugality), rather than an indicator of the absolute amount of money people have, as income was not a significant predictor.

Analysis 2 used a more stringent significance criteria than Analysis 1. This, and the conceptual change in question from ‘who buys more plastic bags at Boots’ to ‘who buys plastic bags 25% or more of the trips when they have bought enough items/volume of items to warrant needing one?’, meant that there were some differences in the variables which were significant. As predicted, in Analysis 2 infrequent single-use plastic bag buyers were more frugal and spent less money in store, suggesting shoppers are motivated to go bag-free to save money. Notably, concern over climate change did not predict single-use plastic bag purchasing - suggesting concern for the environment was not motivating the behaviour. It was additionally anticipated that a lack of self-control or heightened impulsiveness would play a role in people forgetting to take personal bags to the store - and hence would positively predict single use plastic bag purchasing. This was not the case, but high O did. , the results highlight the differences between predicting environmental concerns and predicting a specific environmental behaviour in the real world.

Compared with Analysis 1, BAS variables were no longer significant, and neither was the question on consumer environmental concerns. Nonetheless, the most predictive variables from Analysis 1 remained significant, indicating stability and robustness of those findings. Both models found that age (negatively), average monthly spend (positively), frugality (negatively), and openness (positively) predicted frequent plastic bag buyers at Boots.

Demographic Predictors

The way in which the DV was engineered in Analysis 2 compared with Analysis 1, meant that the prediction accuracy was a lot less, as the model was capturing a lot less of the variance from shopping frequency. This is also reflected in the variables used in the regression. In Analysis 2, none of the shopping frequency variables are significant predictors of bag buying. In addition, age not amount spent is the most important predictor.

That age was such an important predictor in both A1 and A2, where younger people

are more likely to be frequent bag buyers, replicates Poortinga et al. (2016)’s findings. This is contrary to the literature that finds younger people care more about the environment (Hines et al., 1987), but in line with the findings that suggest younger people are less likely to participate in pro-environmental behaviours (Arcury and Christianson, 1993; Honnold, 1984; Klineberg et al., 1998). It is suspected that people who are older are better at planning ahead and bringing personal bags to the store. Increased focus on targeting young people with campaigns or interventions may help to reduce the plastic bags bought. Targeting children and young adults can also help to ensure longer term societal change, as these are our future actors. Further, children can be important social influences on their peers, parents, and community (Hartley et al., 2015).

Following the results from Analysis 1, and the prediction that people would be motivated to not buy plastic bags for economic reasons, it was expected that higher income would predict frequent plastic bag buying. This wasn’t the case. No income category was predictive. Therefore, buying more, no matter what your income, predicts frequent bag buying behaviour. This suggests the mechanism behind bag buying may be more dispositional (i.e. wanting to be frugal, or less materialistic/consumerist) than circumstantial (needing to save money). This may change however, as the government plan to increase the levy in price (DEFRA, 2018a).

Environmental Concern

The ecological motivations for bag buying behaviour were evaluated using two questions: one on consumer environmental considerations, and the other measuring climate change concern. Contrary to hypotheses neither questionnaire item related to bag buying behaviour. This is also slightly different to Analysis 1, which found that climate change concern was not significant, but environmental consumer considerations were. These results indicate that bag buying behaviour has no meaningful relationship with concerns about the environment or environmental considerations when shopping, which is quite unexpected. In research by Sun et al. (2017) it was stated as “obvious” that environmental concern predicts willingness to use plastic bags. However, the findings from this study suggest that the relationship between environmental concerns and plastic bag consumption is substantially more complex.

The questionnaire items used in this study to measure environmental concerns were not selected for the sole purpose of studying plastic bag purchasing, and thus may not fully capture what it means to care about the environment or not. In spite of this, it is prominent that general climate change concern and a pro-environmental behaviour (reducing plastic bag waste) show no significant relationship. This particularly highlights the dissociation between measuring environmental concerns and actual pro-environmental behaviour in the real world. Education on the relationship between plastic bag consumption and climate change could help reduce the number of plastic bags being bought - as those concerned about global warming will be more aware of the detrimental effects, subsequently fuelling motivation to stop buying plastic bags.

Psychological Predictors

As hypothesised, frugality negatively predicted bag buying behaviour - reflecting economic and ecological motivations. This replicates the findings from Analysis 1. Frugality, often seen as the antithesis of consumerism and materialism (Jackson, 2012), also reflects the values of an older generation/past-era of “make do and mend”, pre-dating the throw-away culture of today’s society. Honnold (1984) demonstrated that the mismatch between younger people showing more environmental concerns and older people partaking more in pro-environmental activities was due to era differences in society. These findings suggests that placing more value on thriftiness once again in society, particularly for the younger generations, could improve adherence to plastic bag reduction strategies.

Contrary to predictions none of the variables depicting an impulsive nature: BAS, shopping impulsivity, future focus, or self control were significant in the prediction model. In Analysis 1, BAS-RR was borderline, and BAS-FS switched from having a direct positive correlation with bag buying to predicting it negatively in the regression model - both indicators of weak effects. In analysis two with the additional FDR control, and additionally controlling for shopping frequency by re-engineering the DV, none of these relationships held. This could be because impulsivity is being masked by age, or it could just be that impulsivity isn’t meaningfully related to frequent bag buying.

As in Analysis 1, the only personality trait that was statistically significant when FDR were control for was Openness. The literature found that traits O, C, A, and N significantly predict environmental concern and engagement (Milfont and Sibley, 2012; Hirsh, 2010). It is suspected that because a real world behaviour was measured, and this is somewhat removed from self-reported behaviours and environmental concern/engagement (Kormos and Gifford, 2014), Variant findings were found on the Big Five’s relationship to bag buying behaviour to in the literature. Openness, however appears to be a robust positive predictor of frequent bag buying in transaction data. It is suspected that the facets of openness to experiences most predictive of bag buying are those tapping into impulsiveness, rather than intellect and liberalism (McCrae, 1993). This is supported by the positive correlation between O and BAS in the data (see Figure 6.3). However, facet level O was not measured, and recognise that there are limitations in resolution that comes with using the Ten Item Personality Inventory (TIPI) (Gosling et al., 2003) to measure the Big Five. Therefore this is highlighted as an area which could benefit from further investigation. Despite this limitation, the results suggest that strategies for reducing plastic bag purchasing could benefit from making reduced consumption novel, exciting, cutting-edge, and of great importance - appealing to high O shoppers.

An unexpected finding was that frequent bag buyers rated themselves as being less pleased with how they look compared with infrequent bag buyers. In the data, people who are not happy with how they look have lower well-being, lower self-control, and are also more neurotic (see the correlations in Figure 6.3). Interestingly, this single question about appearance

appears to capture a broader sense of dissatisfaction with oneself and a lack of self-efficacy (which has been shown to be important for engaging in environmental behaviours (Van der Linden et al., 2015; Axon, 2017)). Further research could look further into the effect image dissatisfaction has on other pro-environmental and pro-social behaviours.

6.5 Joint Discussion

Until now very little has been known about the type of people who regularly buy plastic bags. This study used a mass transaction dataset linked to a personality questionnaire to profile the psychological and demographic predictors of regular plastic bag buying. Two different analyses were used to try to better understand the psychological and socio-demographic predictors of single use plastic bag buying. Analysis 1, predicted the frequency of bag purchases over the 9 months, and was much more heavily influenced by shopping frequency. Analysis 2's DV was engineered in such a way to remove the effect of shopping frequency on bag buying frequency. As a result, a lot less variance could be explained in Analysis 2. There are some differences in the importance of the predictors of the two DVs, highlighting the importance of correctly engineering variables in big data so that what you think you are measuring is what you actually are. Nonetheless, some results are consistent across both analyses, suggesting these findings are robust. Both models found that age (negatively), average monthly spend (positively), frugality (negatively), and openness (positively) predict frequent plastic bag buyers at Boots.

6.5.1 Methodological Discussion

Using a mass transaction dataset to study human behaviour allowed the examination of the predictors of a specific ecological behaviour *en masse* in the real world. Using a data driven approach reduced biases from preconceptions and the likelihood of prior knowledge contaminating the results. This approach can be hypothesis generating as well as testing. Many variables were input to mimic, as best as possible, the world we live in. Inputting many variables reduces the likelihood of an unmeasured confounding variable affecting the results, similar to implementing controls in a lab. Real world behaviour, however, is influenced by an infinite number of factors, some of which will be random, and not all can be accounted for. Thus, smaller effect sizes were expected than those that come from a controlled lab-based experiment (Matz et al., 2017).

Big data is an important addition to the study of human behaviour in psychology, but it is not enough to just acquire it. Researchers must utilise the appropriate methods of model validation and evaluation (such as K-fold cross validation and out-of-sample testing), to ensure models do not over-fit. This may require the use of non-traditional psychological methods or statistics (Breiman et al., 2001), but will help to minimise false positives and ensure the findings that emerge are replicable and generalise to a wider population.

Feature Engineering

When working with small data, dependent variables are typically measured either directly (e.g. response times), or indirectly (e.g. using personality scales) where the scales have been formulated and tested. Rarely will psychologists have to engineer the DV themselves using traces of behaviour in data after the fact. In big data research, however, feature engineering is a key part enabling behaviour capture, and transforming the data into a format in which hypothesis can be tested and data modelled. Initially, in Analysis 1, it was believed that the DV was capturing bag buying frequency, and shopping frequency was being input into the model in order to control for the frequency of purchasing. However, it became apparent that if predictions on, not just the behaviour occurring (including the frequency of any shopping behaviour), but on what *psychological* variables predict frequent bag buying (beyond the frequency of purchasing in general) a refinement of the DV would be needed.

Because of the non-Gaussian distribution of the frequency of bag purchases, with the majority of people buying none or one, it was decided to create a discrete binary dependent variable. Indeed, this was done for both analyses, and meant that a cut-off point to define frequent versus infrequent bag buying was required. This could have been chosen in a number of ways. For example, a comparison of very frequent bag buyers to those who did not buy a bag at all could have been made, in order to compare the two extreme ends of the spectrum. Ultimately, the 4 or more purchases cut-off was used for Analyses 1, and the greater than 25% of transactions cut-off for Analysis 2, to maximise the amount of data in the final matched sample of frequent bag buyers and non-frequent bag buyers, whilst still ensuring that there was a sufficient amount variance between the groups.

There were multiple ways the dependent variable could be engineered, and indeed no right or wrong way to do it (although some methods may be more intuitive or common practice than others). The key was to ensure the findings were not an artefact of the way the DV had been engineered. This is where model interpretability is so important - so face validity can be checked for (the DV appears to be measuring what you think it is measuring). For example, in Analysis 1, being able to analyse the β coefficients allowed for the easy identification of the variable shopping frequency explaining much of the variance in bag purchasing behaviour. This led to the re-engineering of the DV to try to delineate the effects of shopping frequency from the psychological and demographic variables.

It is less important how the engineering is done, and more that the engineered variable is interpreted accurately. For example, Analysis 1 could have matched the frequent bag buyers (those who had bought bags 4 or more times) to those who had never bought a bag, in order to reduce the noise around the cut-off (as someone who buys 3 bags isn't going to have a wildly different psychological profile to someone who buys 4 bags), and create a stronger signal in the model. However, then the model would not only be classifying people who buy plastic bags regularly, but also those who never buy them. People who never buy bags may have systematically variant psychological profiles to those who buy bags occasionally. Therefore,

it is crucial that researchers are transparent in their methodology for engineering variables and their reasons behind it, so they can be correctly interpreted and false assumptions are not made.

Masking

In Analysis 1, the logistic regression model was chosen due to it being the most predictive (and simplest) model. However, just because the model performs best doesn't mean it gives the most accurate insights into the relationships between the input features and the dependent variable (Breiman et al., 2001). Despite protecting the model against multicollinearity by removing variables with high VIFS, Table 6.2 shows indications that some variables are still masking others. For example, BAS-RR a variable which it was anticipated would predict regular plastic bag buying, shows a significant correlation with the dependent variables, yet is not a significant predictor in the linear model. Further, BAS-FS is significantly positively correlated with the dependent variable, yet in the regression is a significant negative predictor. This suggests possible interactions or masking of variables. These two examples are particularly pertinent as theoretically BAS-RR and BAS-FS should be related to each other. Even though they are not correlated enough to perhaps affect the overall prediction (using Allison (2012) VIF guidelines) they may be related enough to induce misinterpretation of the regression model.

Although multicollinearity doesn't affect the actual performance of a model, it makes it much harder to determine which variable is responsible for explaining the variance in the DV. Multicollinearity increases the variance of the coefficient estimates making them more sensitive to minor changes in the model. When the coefficients are unstable (i.e. they can easily switch direction) they are much harder to interpret (Minitab, 2013). VIFs are calculated for each variable by performing a linear regression on all other variables, and obtaining the R^2 value. The VIF estimates how much the variance of a coefficient is inflated due to its linear dependence with other factors. A VIF of 1.5 entails that the variance of a particular coefficient is 50% larger than it would be if that predictor was completely uncorrelated with all the other predictors. Unfortunately there is little consensus of how high a VIF can be, before it is too high (suggested values range from 2.5 to 10)(Allison, 2012). Importantly, this analysis has shown that even 2.5 is not sufficient to prevent masking from occurring, or the switching of directions (as with BAS-FS in Analysis 1). Multicollinearity becomes more of an issue when using big data, as the more features a dataset has, the greater the likelihood is that variables will be related. Other methods exist which deal better with multicollinearity, and are exploited on the same data in Chapter 7.

p -values in Big Data

In Analysis 1, a cut-off criteria of $p < 0.05$ was used to determine whether a variable was a significant predictor of frequent plastic bag purchasing; in Analysis 2, a more stringent criteria to control for a false discovery rate of 5% was used. However, in big data, p -values

as an indicator of an effect not occurring by chance, becomes less legitimate the greater the size of the sample. A p -value measures the distance (typically in units of standard deviation) between the data and the null hypothesis using the estimate's standard error. These standard errors shrink as the sample size increases. With a very large sample, the standard error becomes extremely small, so that even the smallest of distances between the estimate and the null hypothesis become statistically significant. This means that in large data, noise or randomness can be found to be significant and this is often incorrectly interpreted as the detection of an effect existing in the general population.

Lin et al. (2013) provides various suggestions of methods to deal with the issue of large p -values in big data in information systems, including the reporting of effect sizes and confidence intervals (as in Chapter 4.3). Other suggestions involve the use of charts and modelling to test hypotheses with varying sizes of samples. However, although these methods provide both researchers and readers a better indication of how the p -values may be affected by sample size, there is no suggestion of how to determine whether an effect is large enough that it is a meaningful finding, given the sample size. Cortina and Landis (2009) identify criteria in which small effects can be considered meaningful. This is discussed further in Chapter 9, Section 9.3. In the present study, that certain variables were significant in both Analysis 1 and 2, across two different samples of different sizes, and with a more stringent p -value criteria controlling for chance, all helps to provide additional weighting to the conclusion that the findings that age, amount spent, frugality, and openness predict frequent bag buying is resultant from sufficient signal in the data. Furthermore, the use of cross validation also prevents noise being modelled ensuring that the overall model is generalisable.

Extracting insights from Big Data

There is great potential for big data to reveal previously undetectable insights about subgroups underrepresented in society or behavioural change over time. However, there is also the danger, if treated the same as small data, of over-fitting models and extracting false or inaccurate insights. There is a difference in extracting variable importance for prediction compared with for explanation (Grömping, 2009). For example, impulsivity may have not been a significant predictor of bag buying (as hypothesised) because of its negative correlation with age, and age's strong positive effect on bag buying. In the regression it is likely that age is masking the effect of its correlates (impulsivity, self-control, shopping impulsivity, A, N) (see Figure 6.3). Depending on the end goal, such variables could be reduced to a latent factor or principle component to improve prediction power (at the loss of granularity in the interpretation) (Harrell Jr et al., 1984); or modelled using Directed Acyclic Graph (DAG), bayesian networks (Nielsen and Jensen, 2009), or structural equation modelling (SEM) to try to unpick the exact causal pathways.

How decisions get made in the real world makes it necessary that policy makers work with a simple and useful model, as they need to trust and understand the information they are basing key decisions on. In such cases finding a few key predictors is often more worthwhile

than modelling a complex network of interrelated features. Thus the zero-order correlations could be useful for practical applications when cost:benefit ratios need to be considered. For example, despite age and frugality being related, by incorporating domain knowledge and judgement it is concluded that age is more useful for targeting individuals than frugality. Age has a greater correlation coefficient with the dependent variable; trait levels are harder to detect within people, yet age is instantly observable; and if one causes the other, it is age that causes increased frugality, not the other way around (Roberts et al., 2006; Specht et al., 2011).

6.5.2 Limitations

This study used a high street health and beauty retailer's transaction data, which is not akin to supermarket transaction data. People may plan ahead by bringing plastic bags for their weekly food shop, but may buy things spontaneously from high street retailers. Or, when shopping on the high street, people may require a bag in the first shop they buy something in, but none of the following stores. Therefore it cannot be ascertained that the results are generalisable to all types of shopping. However, whilst weekly shopping at a supermarket or food store is common for many of us; shopping trips to a health and beauty retailer's are likely to occur less frequently. Therefore, buying four or more plastic bags in 9 months (or more than 25% of times a bag was required) from a health and beauty retailer suggests the behaviour occurs regularly, and will likely reflect a similar type of person who regularly buys plastic bags at supermarkets. However, this would need to be tested before the findings could be confidently generalised to other shoppers.

This research profiled who continued to buy plastic bags after the levy was brought in. As the transaction data for the people collecting free plastic bags from before the levy is unavailable, it can therefore not be investigated for whom the levy changed behaviour the most. For instance, it cannot be ascertained whether people in the income group £25,000-£34,999 stopped buying plastic bags after the levy was brought in to save money, or whether they didn't use as many plastic bags in the first place. Instead, the purpose of this study was to use the data and tools available, to find out who, and how best to target, the regular offenders in terms of plastic bag buying in the post-levy environment.

6.5.3 Implications and Summary

Bans and levies are not the only mechanisms to reduce single use plastics (Xanthos and Walker, 2017) - knowing whom and how to target people is important to avoid a plateau or future increase in bag sales as observed in Ireland (Clarke, 2014). This study has highlighted predictive features of regular bag buyers: namely being younger, spending more, being less frugal, and high in O. Notably environmental concerns were not significant predictors. This profile is dissimilar from that of pro-environmental individuals in the literature and suggests that bag buying behaviour is less about concerns over global warming, and more about the

ability to translate an intention to a behaviour - likely due to the forwards planning required for bringing personal bags to the store (Sun et al., 2017). It is also suspected that the bag buyers are more materialistic and consumerist as they spent more money in store and rate themselves lower in frugality. Whether these effects are caused by age or just correlates is not explicitly tested here, but is unimportant when identifying subgroups such as ‘young people’ to target.

Overall, the drivers of bag buying behaviour measured in transaction data are less characteristic of other environmental behaviours in the literature, measured using self report questionnaires, and more reflective of a consumption-driven behaviour. This highlights the importance of measuring the behaviour, which policy is making decisions on, in the real world. This study has shown that assumptions made in the literature about the predictors of other ecological behaviours did not transfer over to the specific real world behaviour of plastic bag buying. Henceforth, future research should look to harness other digital footprint data in order to provide a more accurate understanding of specific behaviours “in the wild”.

6.6 Conclusion

This chapter used two different definitions of what constitutes a frequent plastic bag buyer. Both Analyses 1 and 2 found that age, amount spent, frugality, and O predicted bag buying. Another prominent finding was that environmental concerns were not a significant predictor of plastic bag consumption. This highlights the disconnect between peoples self-reported concerns about the environment, and their actual purchasing of plastic bags, consolidating the need for data as an objective measure of behaviour in the real world.

As well as identifying the psychological and demographic predictors of plastic bags, this chapter has uncovered methodologies for using both prediction and explanation (which are later formalised in Chapter 9). In both psychology and the field of statistics there is a reluctance to move from data modelling to algorithmic modelling when the size of the sample increases. Using a data modelling approach on big data can lead to overfitting models (by not testing out-of-sample), and inflated goodness of fit measures (due to large amounts of IVS) (Breiman et al., 2001). Further, refraining from evaluating alternative models can lead to inflated type 2 errors (false positives), and ultimately dubious conclusions being drawn. Different samples were used in this data, the smallest being 2,880 people. At what point does the size of the data render a change from a data modelling approach to an algorithmic one? Here, I would suggest, upon visualising the data, when the data is difficult to plot and see structures in (such as interactions, or non-linearity) due to its size or complexity (i.e. the number of data points, features, or interactions). When the data is too big to render it time inefficient to visually investigate its properties, this may be a sign that an algorithmic method is more appropriate.

Overall, social and behavioural scientists should be aware of the advantages and limitations of both approaches in different data environments. Researchers with a wider range of

tools at their disposal will enable better solutions to research problems. Secchi (2018) highlights, more urgently, that if those in the behavioural sciences don't adapt their statistical approaches for big data, there is a risk parametric statistics will "move from fitting models to data to fitting data to models".

Chapter 7

Alternative Methods to Understand Feature Importance for Plastic Bag Consumption

In the previous Chapter, Analysis 1 unveiled differences between the Pearson r coefficients and the β coefficients in the regression model predicting frequent bag buyers. Consequently, it was hypothesised that age could be masking the effect of impulsivity (BAS) due to multicollinearity. A Structural Equation Model (SEM) was used to study the hypothesised interaction between age, impulsivity, frugality, income, and bag buying with the sub facets of BAS combined. An alternative method to investigate the importance of a particular variable, if masking is suspected, is to compare the feature importance rankings of models which are not as severely affected by multicollinearity as β coefficients are in regression. Methods such as decision trees (or random forests) and NMF deal with multicollinearity in different ways, both work well on high dimensional data. This chapter investigates how these alternative models can help supplement variable importance insights from regression when multicollinearity is suspected to be interfering with the results.

7.1 Feature Importance Insights from Decision Trees and Random Forests

In a Decision Tree (DT) or a Random Forest (RF), only one variable is being used to split the data at a time, ensuring two variables are never directly interacting on the same sub-sample of data at the same time. Yet methods of calculating variable importance using decision trees can still be affected by multicollinearity at the point in which the variables are selected to split the data. To illustrate this, let's take the extreme situation of two variables X and Z being 100% correlated. In a decision tree, by chance X will be selected roughly 50% of the time, and Z will be selected roughly 50% of the time. Suppose both X and Z could explain 100% of the variance in the dependent variable Y: rather than the importance being estimated as 100%, either one variable or the other could be selected at each splitting point,

entailing both variables sharing the amount of error they can reduce, exhibiting roughly 50% of the importance metric each. This occurs with traditional permutation importance or error reduction metrics such as reduction in the sum of the squared error (SSE), information gain, and the Gini Index. Likewise with conditional permutation importance, if decisions could be equally made by X or Z, then the predictive power of X is the amount of additional predictive power X brings that cannot be explained by Z. Therefore, in cases where X had an exact correspondence with Y the predictive power of X would be 0% (Strobel, 2008). (This effect is mitigated slightly in RFs due to the selection of the pool of variables being constrained to around 30% of the total number of predictors in order to introduce more variance between the trees.)

Newer methods implemented in the *rpart* package in R (Atkinson and Therneau, 2000) can prevent these issues by utilising surrogate variables to calculate variable importance. CART algorithms (Breiman et al., 1984) can keep track of surrogate splitters (the variable that the algorithm would split on if it couldn't use the primary splitter) as they are useful for when there is missing data. In instances where data from the primary splitter is missing the surrogate splitter is used to help move data down the tree to the daughter nodes. The number of surrogates considered can be entered as a parameter value in the algorithm, with *rpart*'s default being 5. Surrogates can be utilised to help alleviate the multicollinearity problem for variable importance, by allowing the importance measure to include the error reduced when that variable was a surrogate splitter as well as a primary splitter. The surrogates' contribution to variable importance is usually weighted by the strength of the agreement with the primary splitter. The goodness of split measure for surrogates is then multiplied by its agreement score (0-1). Therefore, if two variables are 100% correlated, the agreement score will be 1, and the surrogate and the primary splitter will be attributed equal importance scores. Therefore, rather than [X,Z] being attributed variable importance scores of [0,1],[1,0], or [0.5,0.5] when both variables can explain 100% of Y, both variables are attributed 1 ([1,1]).

Using decision trees and RFs with variable importance calculated in this manner, the importance rankings of features can be compared to those in a logistic regression, to better understand if masking was occurring. For example, if impulsivity is an important feature in a decision tree and random forest, but it isn't in the logistic regression, one may hypothesise that it is being masked by other variables in the regression model. Alternatively, it could be that impulsivity has a non-linear relationship with plastic bag consumption, but not a linear one. However, if it is not an important feature in the decision tree or random forest it is likely that it was truly not a good predictor.

7.1.1 Decision Tree Analysis

The decision tree in this analysis was built using the data from Analysis 1 in Section 6.2. Firstly, the decision tree's meta-parameters were optimised using 5-fold Cross Validation

(CV) on 80% of the data. Different parameters were iterated through using a grid search technique. The parameters were evaluated using classification accuracy (the fraction of predictions the model got right). Once the optimal parameters were selected¹ the optimised tree was tested on the 20% of data left out. The decision tree predicted the test cases with 73% accuracy, 3% less than the logistic regression. The confusion matrix can be found in Figure 7.6c. The results of the full comparisons to other models in this section can be found in Table 7.2.

To extract the variable importance the tree was fitted to 100% of the data, keeping the optimised parameters. Variable importance was calculated using the reduction in node purity, using five surrogate variables. This was converted to a percentage (of the total amount of impurity reduced by the tree) to enable later comparisons to the RF variable importance rankings. Note that when using incremental node purity, as no significance or effect size is being estimated, there is no need to control for false discovery rate or inflated coefficients. The decision tree, fitted to 100% of the data, consisting of 10 binary splits, and is shown in Figure 7.1. The variable importance extracted from the tree's primary and surrogate splits can be found in Figure 7.2.

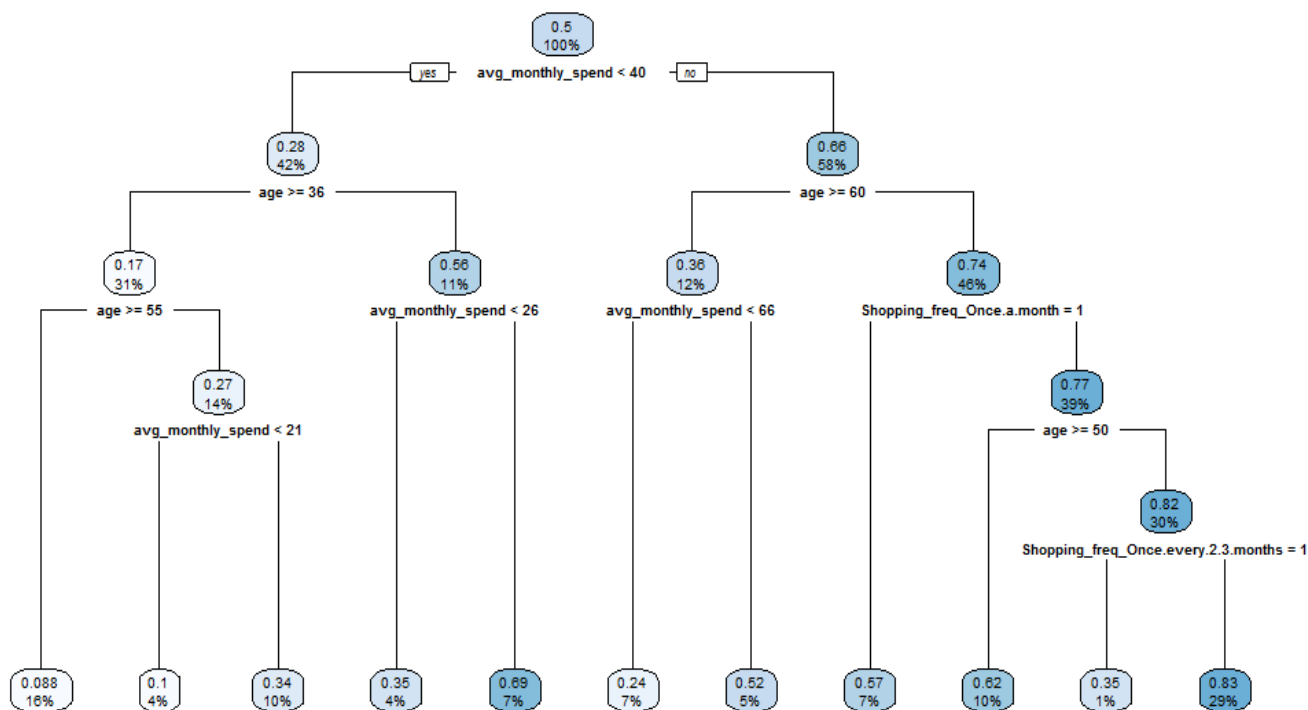


Figure 7.1: Decision Tree model for the optimised tree fitted to 100% of the data, pruned at the point it at which the cross validated error is lowest. For the dependent variable predictions at the terminal ‘leaf’ nodes along the bottom, a 0 is a non-bag buyer, and a 1 is a regular bag buyer. Spending is in pounds (£). For the shopping frequency questions 1 = ‘yes’, 0 = ‘no’.

¹Optimal meta-parameters for the decision tree were: minsplit = 2, minbucket = 2, cp = 0.01, maxdepth = 30, usesurrogate = 0, surrogatestyle = 0

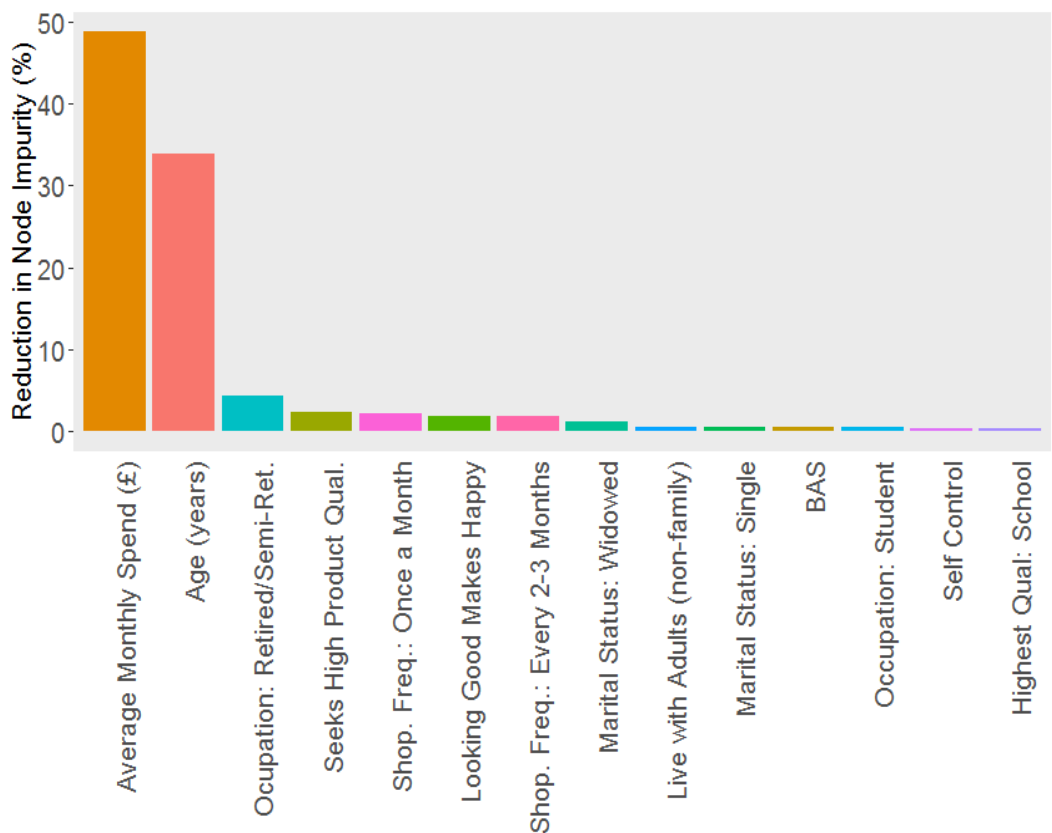


Figure 7.2: Variable importance (calculated as the % decrease in node impurity) in the Decision Tree model for the optimised tree fitted to 100% of the data, pruned at the point at which the cross validated error is lowest.

7.1.2 Insights from the Tree

The decision tree, shown in Figure 7.1, uses just three features in 10 binary splits in order to perform at an accuracy 3% less than that of the logistic regression (where 16 variables were significantly contributed towards the prediction). The three variables used in the decision tree model are average monthly spend, age and self-report shopping frequency. The tree tells us that spending more on average a month than £40 is the greatest predictor of regular plastic bag buying. After that, age is used to divide the sample into 4 subgroups where those who are most likely to be regular bag buyers (0.74 probability) spend >£40 on average a month, and are younger than 60 years old (46% of the sample); and those who are least likely to be regular bag buyers (0.17 probability) spend <£40 on average per month, and are older than 36 years (31% of the sample).

The first two times in which age was used in the tree, the data was not split using the mean age values (51 years), suggesting that the relationship may be non-linear or there may be an interaction with the amount spent per month. These possibility are analysed in Figure 7.3. The graph shows an interaction between age and amount spent per month (using a mean cut off of £40) creating 2 subgroups: one is older and spends more, the other is younger on average and spends less. Combining these two variables could improve the prediction of frequent bag buying. To explore this, age and amount spent were input in to a simple linear regression model, along with their interaction, to predict frequent plastic bag buying. Not only was the interaction significant ($t(2876)=-3.99, p<0.001$) but the whole model could explain 24% of the variance in bag buying using just the two features (spending

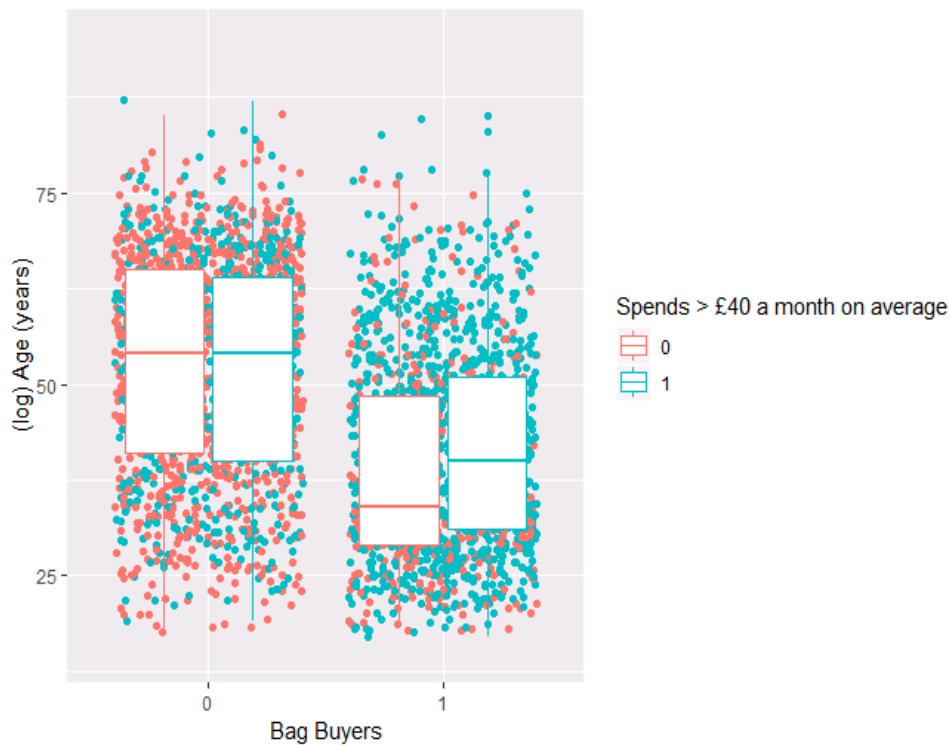


Figure 7.3: A graph to illustrate the interaction between age and average monthly spending when predicting frequent plastic bag buyers

and age) and their interaction ($R_2=0.24$, $F(3,2876)=306.9$, $p<0.001$).

Overall this analysis suggests, that when the variables age and those about shopping regularity and amount spent are measured, that psychological variables are not useful to predict plastic bag consumption. This doesn't mean to say that when age isn't measured, impulsivity and other psychological variables like self control cannot be used. To illustrate how these variables could substitute in, the variable importance comparisons in Figure 7.2 show that, despite BAS and self control not appearing in the tree structure used, they were used as surrogate variables (as they have some importance attributed to them). When the surrogate lists were printed out, these confirmed that BAS and self control were used as surrogate variables for age.

7.2 Feature Importance insights from a Random Forest

For a decision tree, small changes in the data can cause large structural changes to the tree structure, and thus potentially the rankings of the important features. Whilst, a random forest is a more stable model, and thus the feature importance rankings are more robust (Breiman, 2001). Therefore, generally speaking, the insights gained from a random forest's variable importance rankings are more generalisable. Additionally, the RF was a better prediction model than the decision tree (see Table 7.2) suggesting that its solution (and the combination of variables it used) is more "correct". Hence the variable importance from the random forest was also analysed, using the same method as for the decision tree, and can be viewed in Figure 7.4.

Like the decision tree, the two most important predictors of plastic bag buying were average monthly spend, followed by age. Interestingly though, the RF found frugality, followed

by shopping impulsivity more useful than any other of the demographics variables, shopping motivations, or shopping frequency. This is in line with the results of the logistic regression. The next most important variable, interestingly, was duration, which was removed in the feature selection phase before being input into the logistic regression (see Section 6.3.1. This highlights that there is most likely only a non-linear relationship between the time it took someone to complete a questionnaire and the number of plastic bags they bought. This could be tapping into some shared latent psychological trait or thought pattern.

Because multicollinearity does not affect the interpretation of decision trees and random forest’s variable importance as much (when calculated using surrogates), the BAS variable overall (as well as its sub facets) were input into the model to better understand *if* impulsivity is an important predictor, *which* parts of BAS (or BAS as a whole) are related to bag buying behaviour. The variable importance results from the RF show that self control (the opposite of impulsivity) and BAS overall were the 7th and 8th most important variables respectively. Drive (BAS-D) followed, and then Reward Responsiveness (BAS-RR) and Fun Seeking (BAS-FS) was the least important impulsivity facet. Interestingly, all of the Big Five personality traits, except conscientiousness, were ranked as more important by the RF than both environmental concern variables (ranked 20th and 21st respectively).

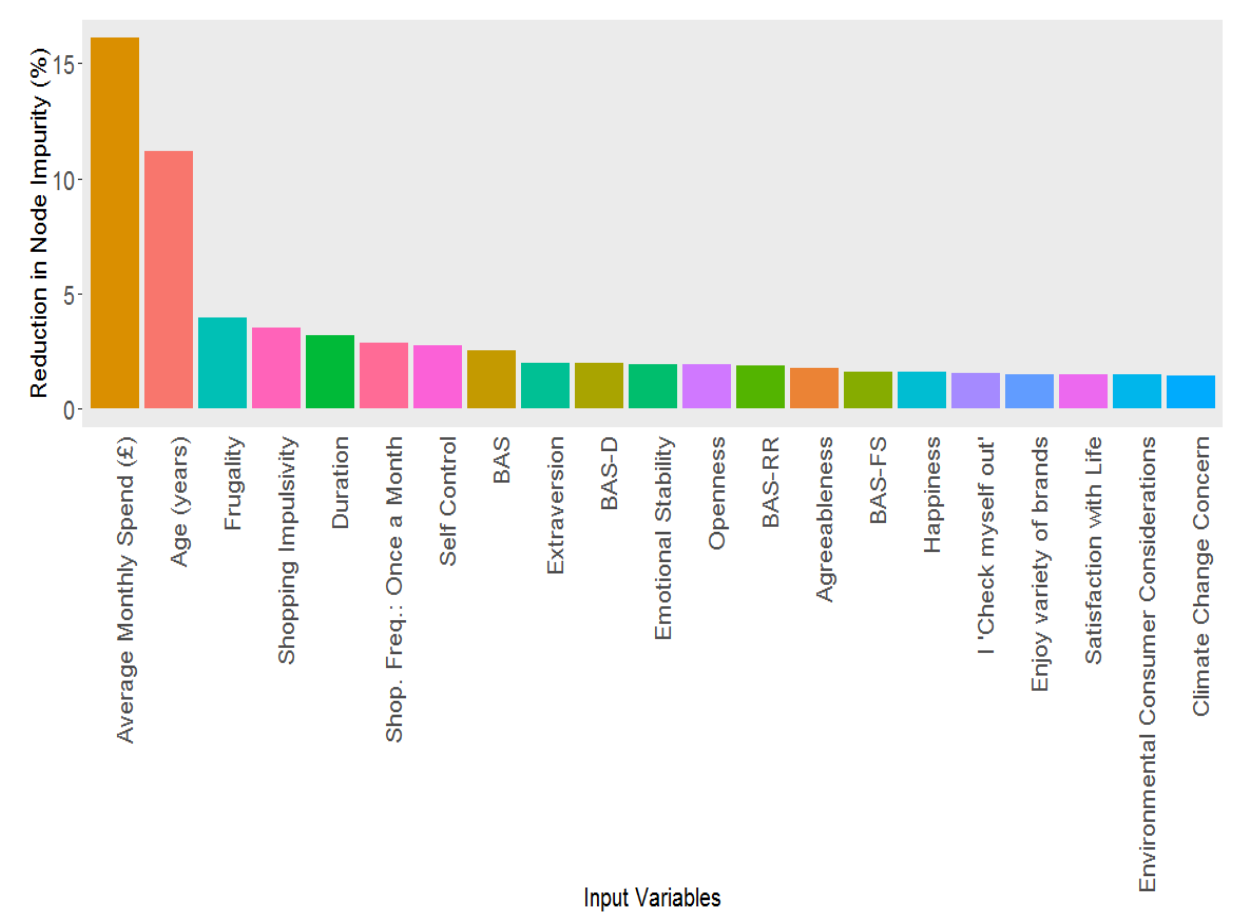


Figure 7.4: A graph to illustrate the percentage reduction in node impurity for each variable in the random forest

7.3 Feature Importance Insights from NMF

Nonnegative Matrix Factorization (NMF) is a matrix factorization method used for feature extraction or reduction, where the matrices are constrained to be nonnegative. It provides a similar output to principle components analysis (PCA), except that as coefficients (loadings) are restricted to be positive, it is easier to interpret the factors that emerge. NMF reduces a feature space into two separate matrices, which when multiplied together, retain the information of the original feature set. The factors which emerge, that the features are combined to load onto, are called *bases* (singular, *basis*) (the extent of the loading is called a coefficient).

NMF is useful when there are many features which are ambiguous or have weak predictability. By combining features, NMF can reveal topics and themes in the data. By looking at themes as combinations of measured variables, NMF can also help to identify potential latent variables or structures in the data. NMF is useful for wide data, or when multicollinearity occurs. If two features are highly correlated they will likely load onto the same factor, being combined to predict the outcome variable. If the factor they load on to is predictive, the interpretation is that they are both useful (rather than one or the other). If one variable loads more heavily onto the predictive factor, it is interpreted that this variable contributes more to the prediction. Extracting the correct number of basis is a balance between retaining a good amount of information, and ensuring the bases are orthogonal to one another (don't have shared variance).

7.3.1 NMF Analysis

Before conducting the NMF analysis, features were first normalised to ensure all values were between 0 and 1. Thereafter, different numbers of bases were iterated through to calculate the residual square error for each basis solution (see Figure 7.5). The variable loading's on different basis solutions were also viewed and analysed to see if conceptually there was an optimal solution. A solution of ten bases was decided on, as when the (loading) coefficients were input into a regression model, it predicted plastic bag buying the most accurately in training. The ten bases and their interpretations can be viewed at the end of this Chapter (Figure 7.5).

Once a 10 basis solution was chosen, the coefficients were extracted and input into a logistic regression model. The regression was subsequently optimised on training data, using a grid search and 5-fold validation. The optimal solution gave an accuracy rating 66% on training data. The model was then tested on a 20% sample of unseen data, for which it barely performed better than chance (see 7.2). The confusion matrix in 7.6d shows that the NMF predicted many false positives - it predicted many people to be regular bag buyers when they were not. This highlights the importance of using different metrics to evaluate model performance, not just relying, for example, only on the total of cases accurately predicted (particularly if there is a cost to wrongly predicting positives). Furthermore, the estimates

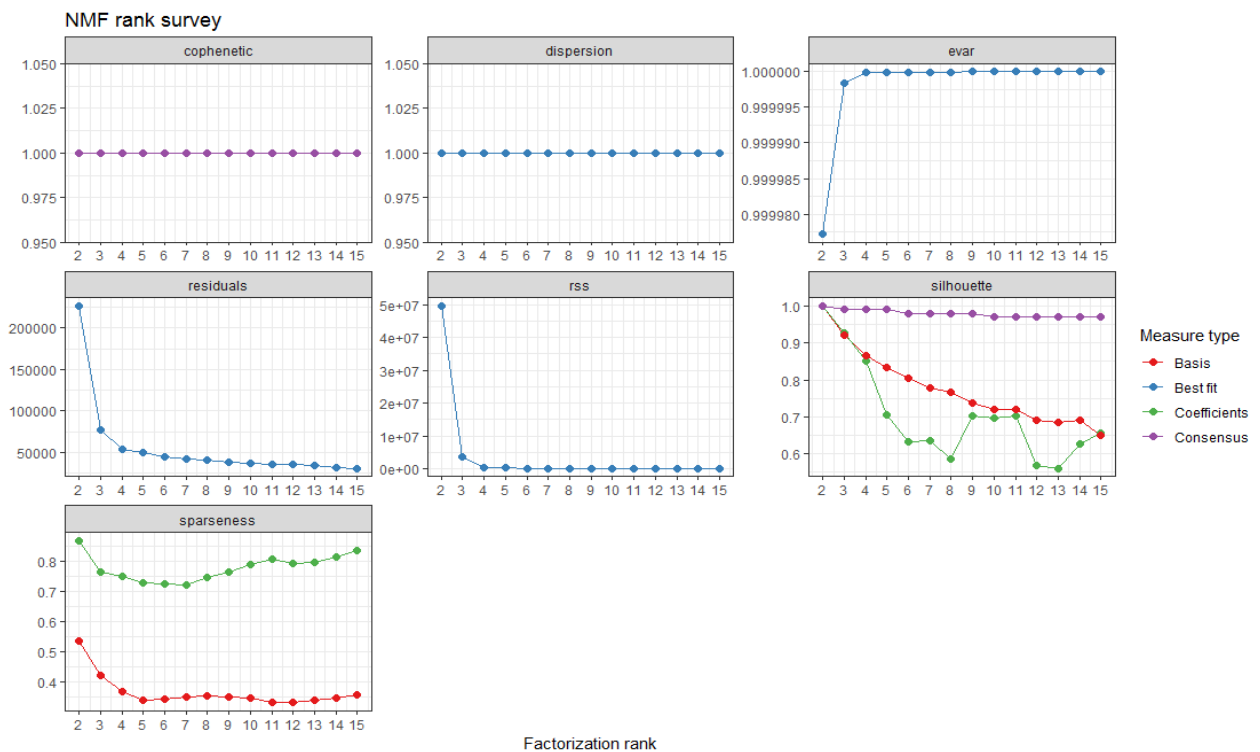


Figure 7.5: Plots to illustrate the evaluation metrics used to choose the number of basis in the NMF. The RSS (residual sum of squares) graph illustrates how the majority of the variance was captured in the first four bases.

of model performance on the training data were much better than on out-of-sample data, suggesting that for some of the validation samples the model was overfitting. This highlights the importance of, at the very least, averaging model accuracy over cross validation folds on training data (one must never assume that the accuracy achieved from the fold with the optimal solution is a valid metric), or ideally out-of-sample testing.

Despite the poor performance on test data, when the model was fitted to 100% of the data, 9 out of the 10 variables were significant predictors of bag buying behaviour (only basis 3 was not, see table 7.1). The model could explain 12% of the variance explained ($R_2=0.12$, $F(10,2869)=42.35$, $p<0.001$). This illustrates that even a poor model, where the predictions are only 1% better than chance, can still yield significant results. This example illustrates that a significant effect doesn't always correspond to a meaningful effect. This is particularly clear when these results are contrasted to the regression presented earlier using just three variables (age, average spending, and the interaction between the two), yet explaining double the variance in bag buying behaviour.

Table 7.1: NMF regression results

	B Coeff	Std. Error	t value	p sig.
Basis 1	0.59	0.29	2.03	*
Basis 2	0.28	0.08	3.35	***
Basis 3	-0.15	0.08	-1.94	
Basis 4	1.27	0.13	9.62	***
Basis 5	0.65	0.11	5.94	***
Basis 6	0.79	0.12	6.67	***
Basis 7	-0.59	0.09	-6.85	***
Basis 8	-0.98	0.13	-7.60	***
Basis 9	0.46	0.08	5.49	***
Basis 10	-0.26	0.13	-2.03	*

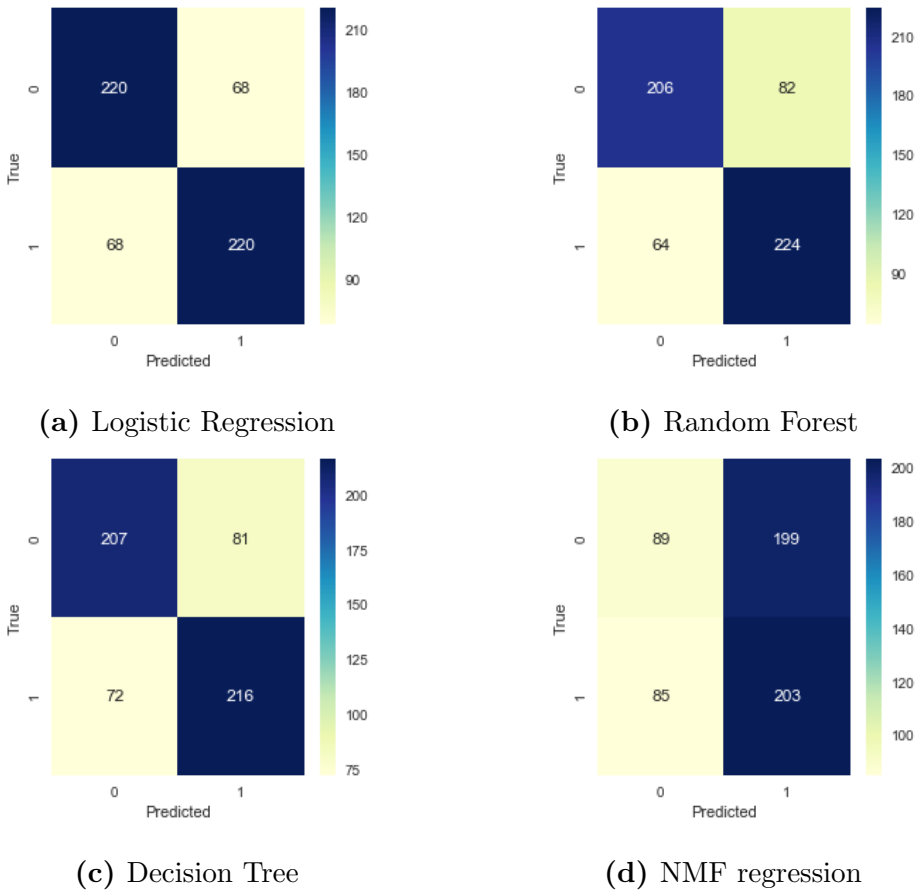


Figure 7.6: Confusion matrices for a) the decision tree and b) the NMF regression models on the test data. 0 = non bag buyers, 1 = regular bag buyers.

Table 7.2: Model Comparisons

	Number of features	F1 score	Accuracy	Precision	Recall
Logistic regression	84	0.76	76%	0.76	0.76
Random Forest	95	0.75	75%	0.73	0.78
Decision Tree	95	0.73	73%	0.73	0.73
NMF regression	10	0.49	51%	0.51	0.51

7.4 Discussion

In addition to the variable importance extracted from the logistic regression in the previous chapter, this chapter presented additional analyses from three other models. This was to investigate further the predictors of frequent bag buying behaviour using methods where multicollinearity would not affect the interpretation. A random forest found that age, amount spent, and frugality were important predictors. This consolidates the findings from the regression model in Chapter 6, allowing conclusion to be drawn on the robustness of these features as predictors of plastic bag buying.

NMF proved to be distinctly not useful, with the majority of the psychological variables loading onto one feature, and the resulting regression model generalising poorly. Whilst decision trees and random forests enabled the greater insights to be extracted from the data. The structure of the decision tree revealed that it is not just age and amount spent which are the top predictors on their own, but that they can be combined to explain even more variance using their interaction term. The decision tree also showed that other variables such as BAS and being retired can step-in as surrogate variables for age.

The random forest provided insights into the non-linear psychological predictors of frequent plastic bag buying. The random forest found that after age, average spending, frugality, and shopping impulsivity were important predictors of bag consumption. As impulsivity was not predictive in the linear regression model, it can be concluded that shopping impulsivity (impulsivity, as expressed in a consumer context) has a non-linear relationship with plastic bag buying. This does not necessarily mean it fits a non-linear function/interaction, but could be that in certain subgroups, shopping impulsivity is highly predictive (i.e. perhaps within a certain age group). Moreover, self control and BAS were the 7th and 8th most important predictors in the RF. This provides further evidence that some impulsive tendency is responsible for bag buying behaviour, and that atleast some of this is irrespective of age.

7.4.1 Is Statistical Significance Important?

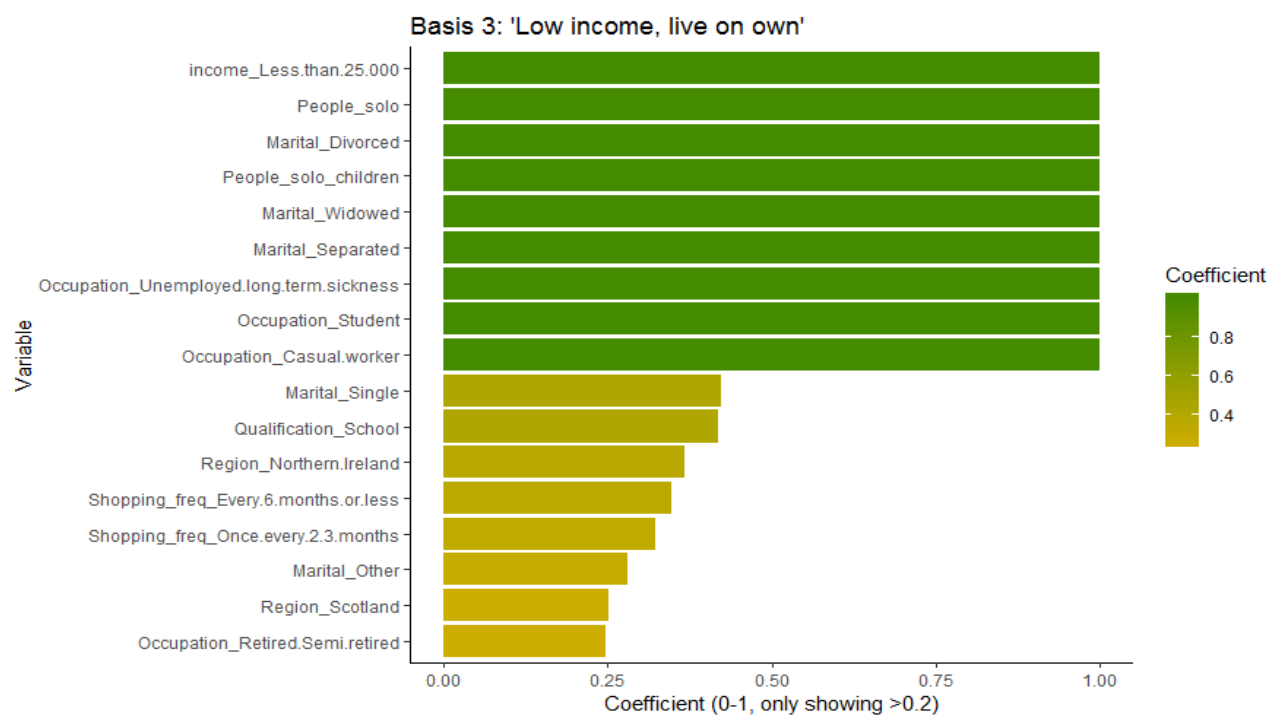
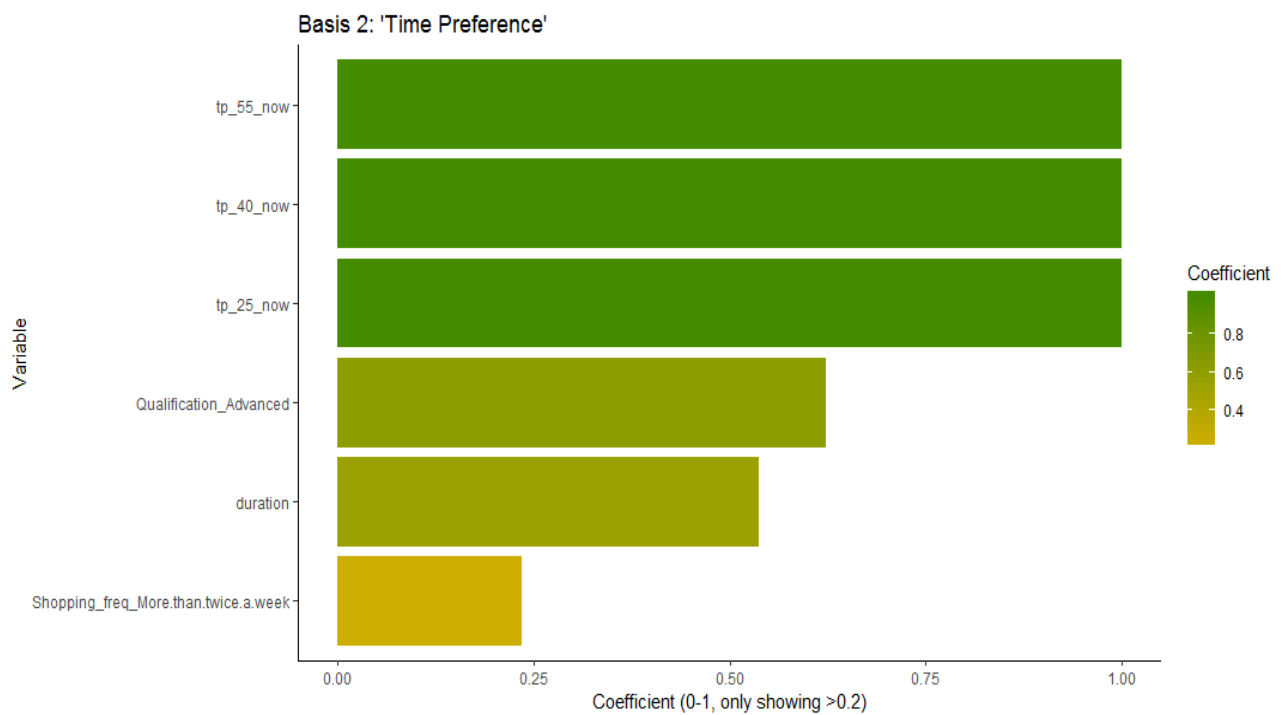
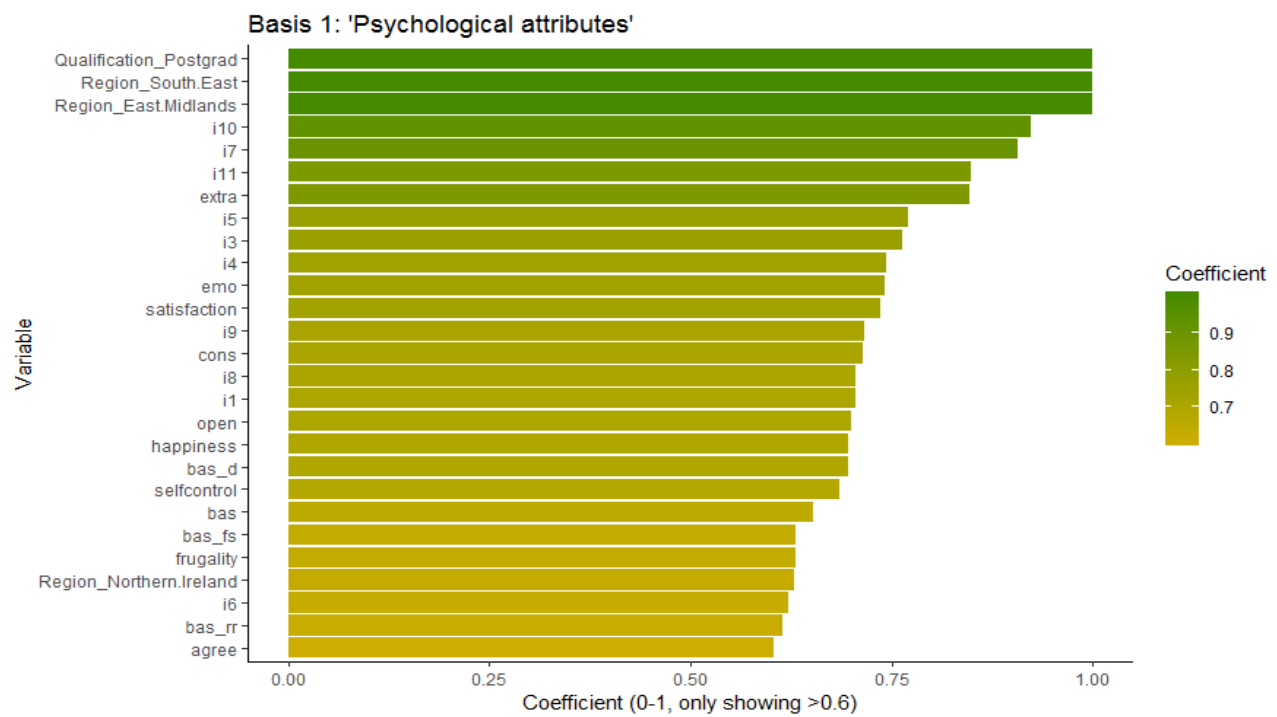
Throughout the investigations into plastic bag buying in this chapter and the last, the logistic regression model was repeatedly the better model, suggesting that predominately linear relationships occur between the features and plastic bag buying. However, if the random forest had been the better model, how much of a drawback is to not be able to identify the features which are statistically significant? When reporting results for a psychology journal, being able to ascertain significance seems to be crucial, yet in the machine learning field it is not.

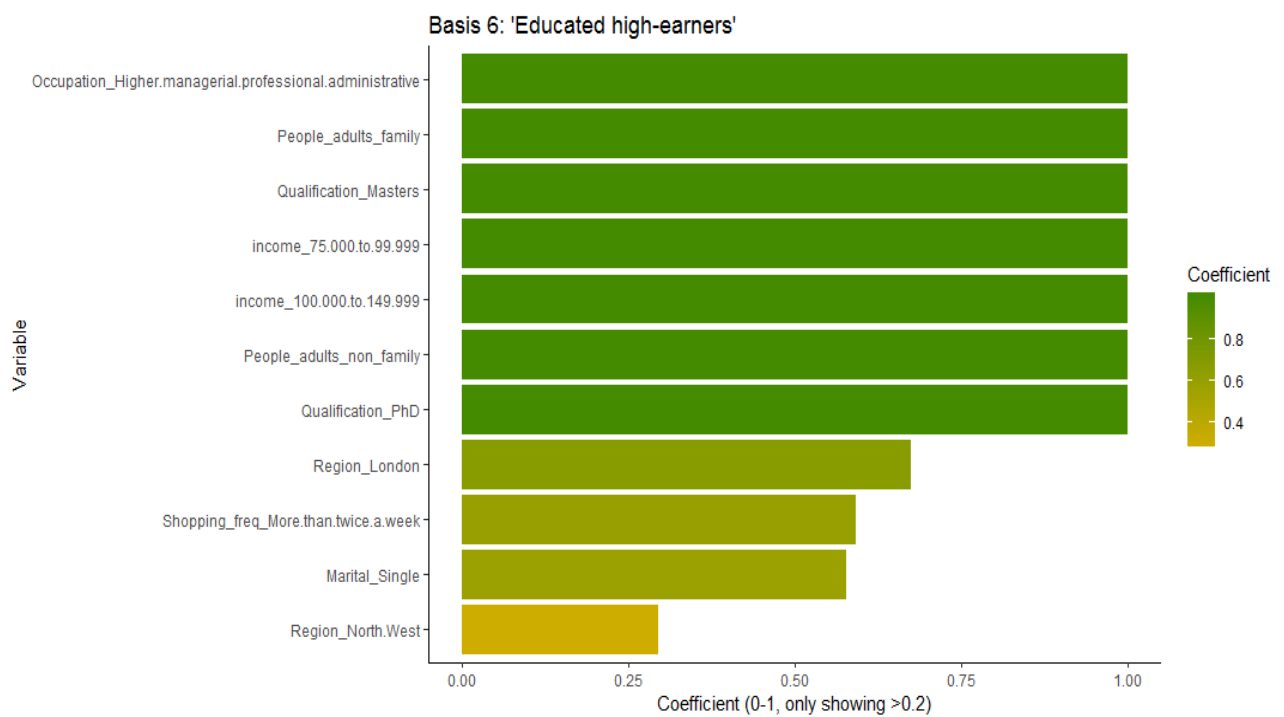
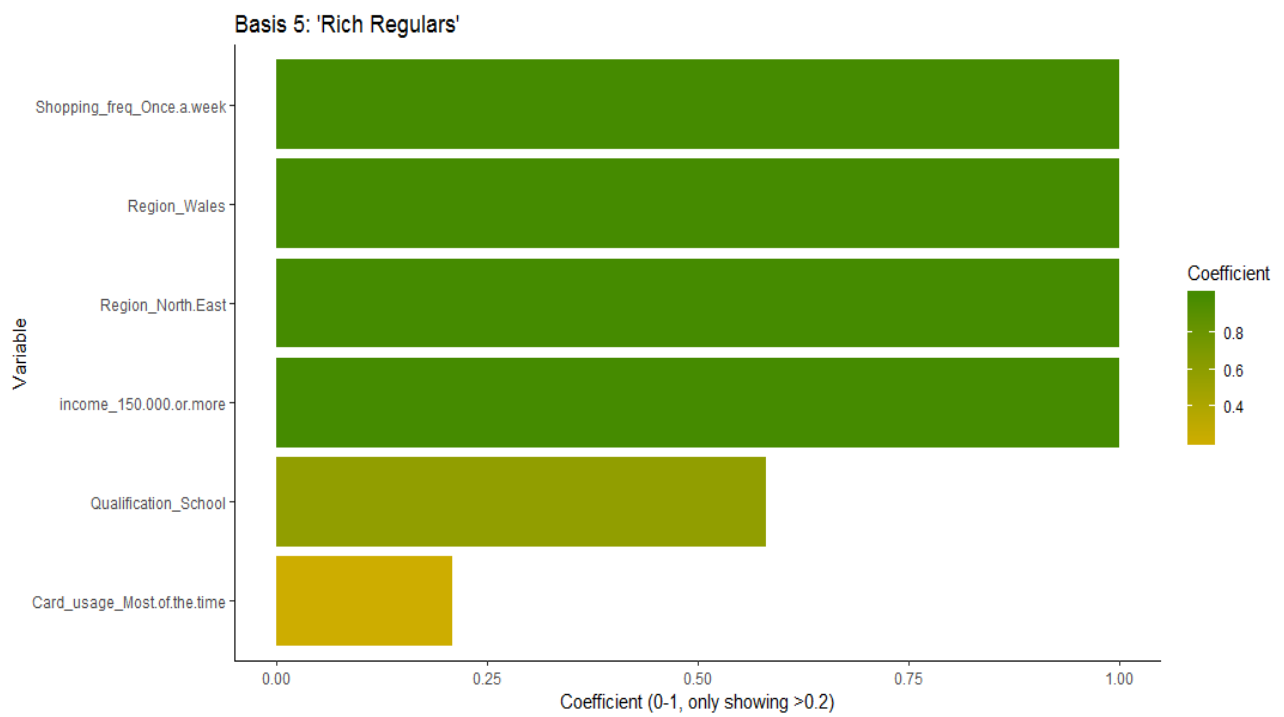
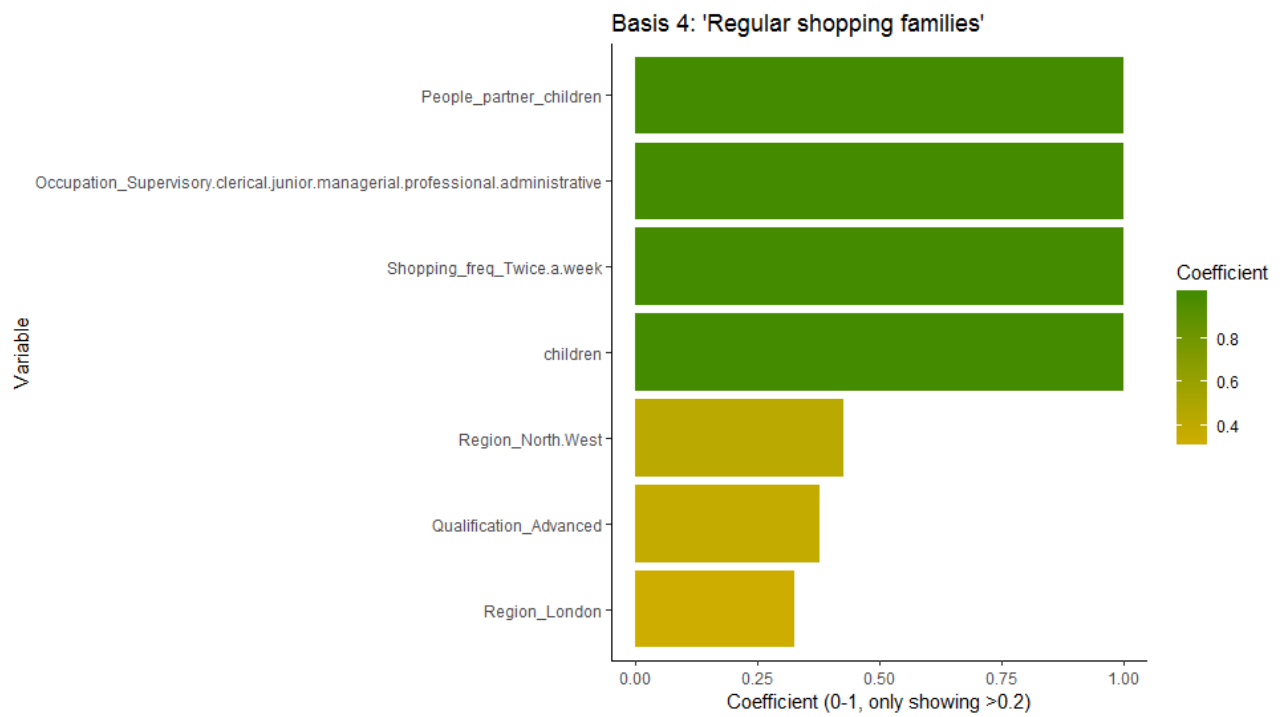
In big data, statistics become less important, as if a sample is large enough, any effect will become significant (see Section 6.5.1). If a model is useful, and variables are important to that model, then isn't this enough? This may be the case for highly predictive variables, such as age, but it is less clear cut for smaller effects (or the variables which only contribute a small amount to the prediction). How do we determine when a variable is providing a meaningful contribution, and what is the cut-off for this? These issues are discussed further in Chapter 9.

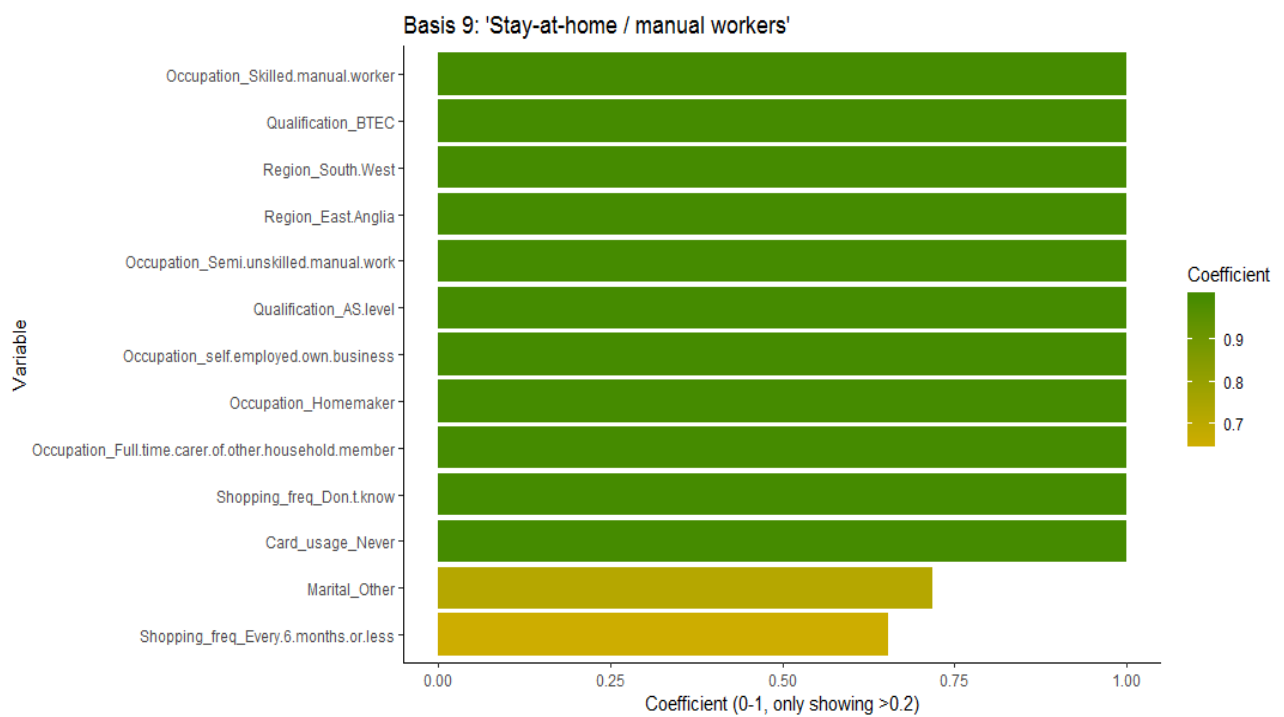
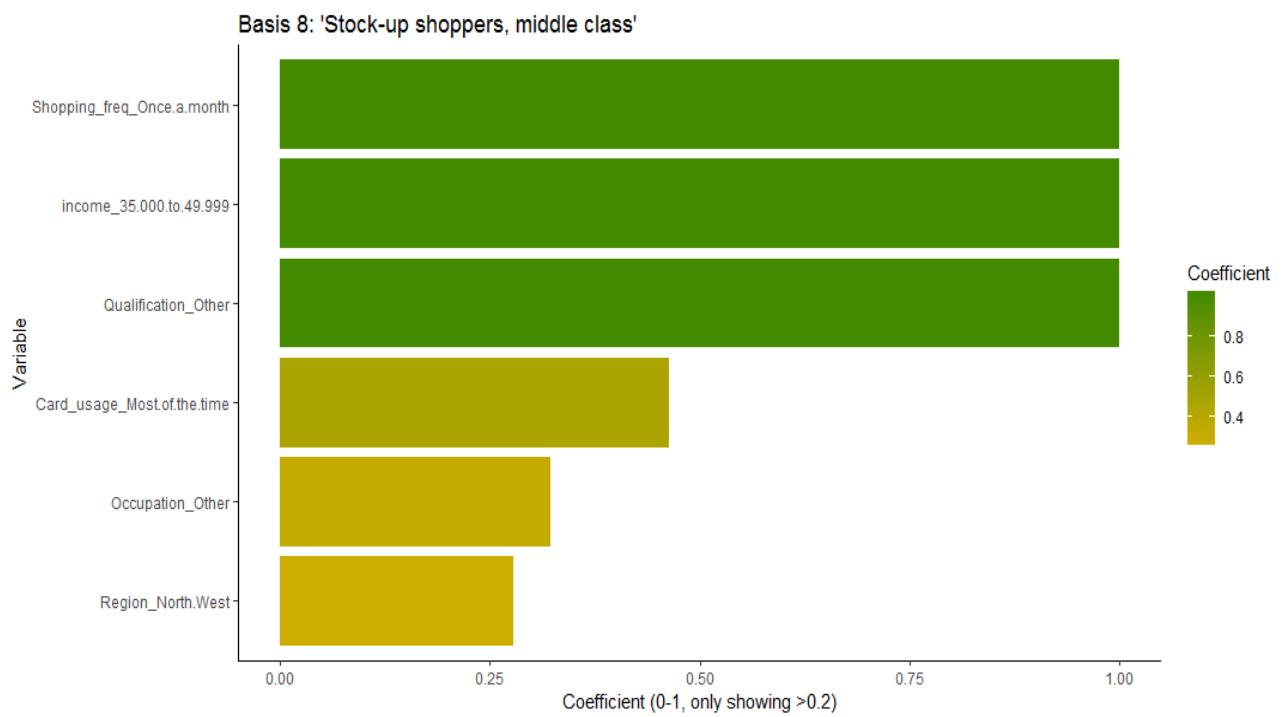
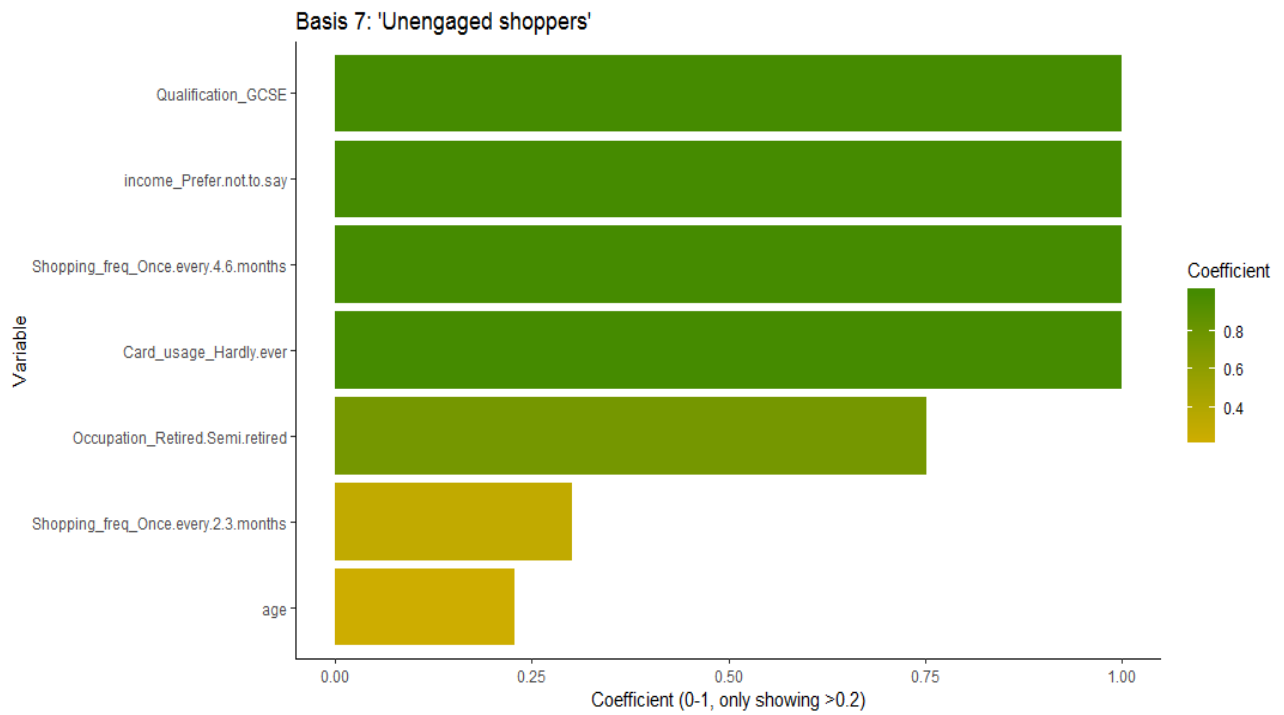
7.5 Conclusion

Taken together, the findings from these three supplementary analyses suggest that impulsivity is predictive of bag buying behaviour, but it is either non-linearly, or within particular sub-groups (most likely created by spending and age given the structure of the decision tree), or both. It is deemed legitimate here to draw such conclusions from the RF, as although it was not the most predictive model, it was only $<1\%$ less accurate than the linear model.

In Chapter 6, Section 6.3, an algorithmic approach was used to test multiple models before selecting a single one. However, in this Chapter it has been shown that additional insights can be gained from other models if the following conditions apply: 1) the model interpretation of the “winning model” is suspected to be affected by multicollinearity, and 2) the “runner(s)-up” are not that far behind (i.e. don't show a large reduction in prediction power). Using models which have different underlying mechanisms can allow the researcher to put different pieces of information together to gain more detailed and nuanced insights about what might actually be going on in the data. This approach, discussed in more detail in Chapter 9, highlights that perhaps researchers shouldn't be as tied to one explanation from one model, as is typical in psychology.







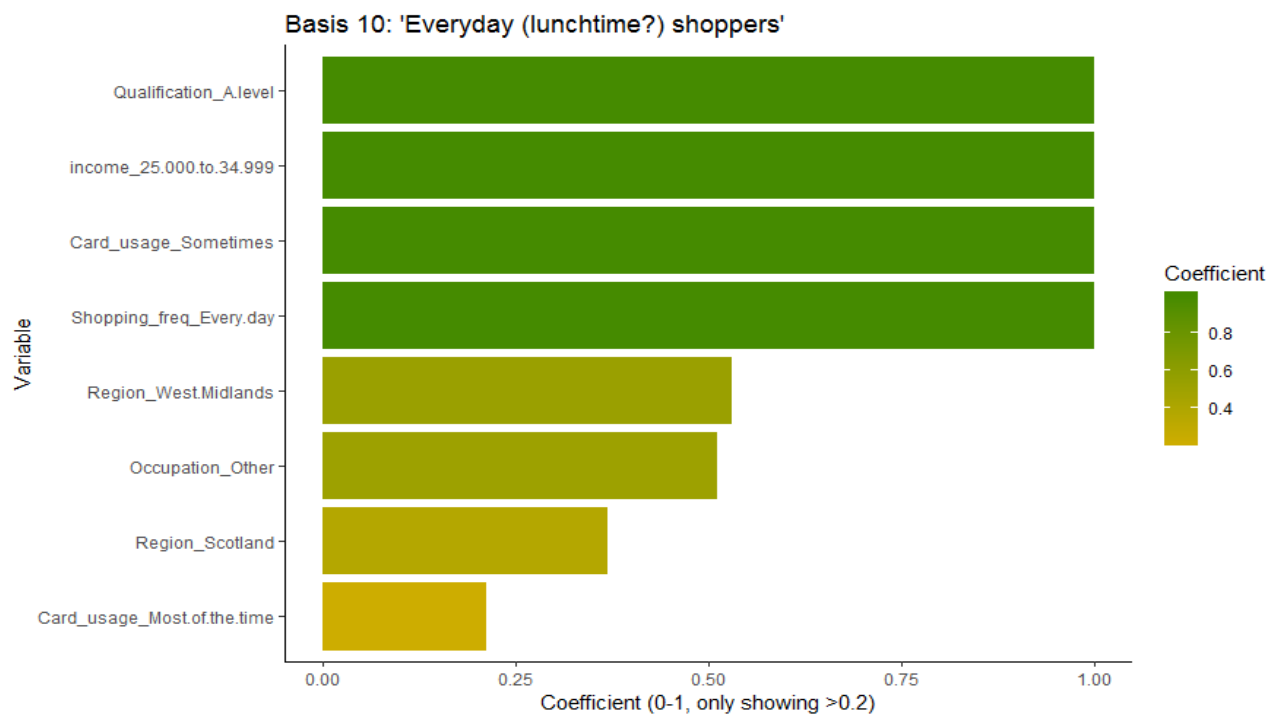


Figure 7.10: Plots to illustrate the loadings of the different variables onto the bases, and how the bases were interpreted (written above the graph).

Chapter 8

Investigating a Behavioural Measure of Anxiety in Transaction Data

One of the proposed advantages of big data is the potential for it to be used to measure behaviour, reducing the reliance on self-report measures of behaviour. But what if big data could also augment self-report personality assessments, or psychiatric assessments? Personality assessments typically require introspection, but what if, as well as being a measure of behaviour in the real world, big data could measure internal psychological phenomena (or at least the behavioural expression)? Personality is one of the drivers of behaviour, where people of a similar personality are more likely to behave the same (Tkalcic et al., 2009), thus certain behaviours may be reliable reflections of certain personality types.

In Chapter 4, the purchasing of different product types were indicators of a particular personality type. For example, the purchasing of smoking cessation products was a strong indicator of an extrovert, emotionally stable personality; whilst incontinence pads were an indicator of introvert agreeable people. These examples are just simple illustrations of how a personality could be expressed through consumer behaviour. More complex algorithms would be able to detect stronger, more reliable signals. If these can be detected, then the expression of personality in behaviour could be measured externally, instead of having to ask people questions. The idea of using behaviour as an indication of personality is not a new phenomenon. Many personality measures involve questions about whether a person enjoys or engages in particular behaviours, as an indicator of their internal trait constructs. For example, sensation seeking (a facet of Extraversion (E)), is often measured by asking respondents to answer how much they agree with the statement ‘I enjoy skiing quickly down mountains’.

This chapter first discusses in more detail how big data could be used in psychology, not just to predict and understand the behaviours themselves, but how the behaviours captured can help to interrogate the psychological constructs researchers use. In this way, teachings could flow both ways between behavioural data and personality theory. We can better understand human behaviour in big data using personality constructs; and we can use the complex behaviours captured in data to assess the content and measurement validity of the personality constructs. For example, with big data encapsulating a rich variety of

dynamic behaviours, if traits can not predict particular behaviour, perhaps this reflects an incompleteness in the constructs psychologists are currently using.

The chapter then offers a preliminary investigation into whether the purchasing of over the counter anxiety medication could be a behavioural measure of anxiety in the general population. The behavioural measure is first tested for face validity by assessing whether it is related to the personality traits E and Neuroticism (N) as expected from the literature. The behaviour of purchasing anxiety medication is then investigated further by predicting it using the self-reported variables from the questionnaire. It is theorised that the purchasing of over-the-counter anxiety items may reflect a slightly variant type of anxiety expression, to that typically measured using self report. The strongest psychological predictor is shopping impulsivity, suggesting that the behaviour may be a consumerist, potentially impulsive, reaction anxiety symptoms. However, before such behavioural measures of anxiety can be realised, prediction accuracy would need to be dramatically improved. Below, the reasons why predicting individuals' behaviour in big data is so hard, are discussed.

Another advantage of big data, is that much of it is geo-tagged (has location information attached to it). The final section of this chapter analyses personality data spatially, to identify whether geographic personality clusters exist in the UK. This is done using where people shop at Boots as a proxy for which district they reside in. The assumption that where people shop at Boots is a proxy for where people live is then interrogated by analysing two other datasets: BBC personality Test data, and the UK census data. Finally, it is investigated whether the personality of a store's customers correlate with the sales in anti-anxiety medication.

8.1 Introduction

Content Validity and Big Data Psychology

Most research using big data and personality focuses on prediction, for example, predicting personality from Facebook profiles (Kosinski et al., 2015), or shopping patterns (Gladstone et al., 2019). Such research uses behaviour to try to estimate an internal underlying trait which has been measured as the ground truth. Much rarer is the reverse process of using data on what is taking place in the real world to better understand personality, and to feed back into the tools used to measure what is going on in someones head. Personality does not exist as its own entity, it is a construct humans have invented in order to better predict how people are going to behave. A better understanding of why the psychological constructs developed are good at predicting behaviour, and where they fall short, can help to develop and improve how psychological traits are conceptualised and measured. Being better able to describe the internal constructs which are responsible for behaviour, feeds back into feature engineering and the design of prediction models, enabling more accurate behavioural prediction.

Personality measures were originally created by asking people to describe people's characteristics and then performing factor analyses using the words (Goldberg, 1990). Conse-

quently, personality scales are limited by language, in their ability to capture information. In the age of big data, as well as linguistic descriptions for personality, a person's behavioural records act as an expression of their personality. Thus, if millions of people's personality traits are acquired and linked to their data, detailed descriptions of individual differences measured in behavioural traces can be attained. From people's digital footprints individual differences can be traced and described through behaviour. These different behavioural patterns that vary with personality could be used to re-assess whether the same constructs emerge in behaviour as with language.

If behavioural alternatives to rating scales can be developed, then big data can also be utilised to construct temporal measures of personality. Measuring personality over time using questionnaires is time inefficient and expensive. If equivalent behavioural measures were created, however, temporal processes could easily be studied as no additional information would need to be collected. Instead, the streams of data already being retrieved and stored could be exploited. This is an area that personality research has not yet thoroughly explored. Using big data to study construct validity could help to answer crucial questions in the field of individual differences. For example, when personality change is detected over a life span, is it how people *express* their personality that changes (i.e. it is an artefact of the measurement tool), or is it the internal structural biological mechanisms that change (Watson, 2004)? A common question on a personality questionnaire to measure extraversion using a likert scale is "rate how strongly you agree with the statement 'I am the life and soul of the party' on a seven point scale from 'strongly disagree' to 'strongly agree'. An older person may not go to parties anymore, and thus answer 'neither agree nor disagree' despite, for a 90 year old, being extremely extroverted. Therefore, this question is not able to capture 90 year old Doris's level of extraversion correctly, as it is not calibrated for someone of her age.

This calibration issue is not specific to age either, it can also occur for different life stages. For example, take the example of Linda, a mum of small children who has forgotten how she behaves at parties as it has been so long since she went to one! And perhaps Linda may have loved parties in her 20's, but having a family changed her priorities and she no longer cares for them now. Has her personality changed or has something else changed? Linda may have replaced parties with some other outlet, something which may be apparent in her data, something which is taken on by other individuals who are highly extrovert and *are* the life and soul of the parties. Through such behavioural associations, a richer and fuller metric for personality can be established. Thus, rather than having one measure of personality, having different behavioural measures of personality which are age and life stage specific, could reveal interesting insights about temporal stability in personality. If personality measures are better calibrated between age and life stages, personality may be more stable over time than current research suggests (Watson, 2004; Roberts et al., 2006; Specht et al., 2011).

The same logic applies to different contexts, as well as different ages. Often people behave differently in different situations or different circles of people. Using big data, this can be quantified and a concept of situational personality expression may be developed. If

context-calibrated personality measures were developed, it should be possible to say whether people's personalities remain stable over contexts or whether people switch between expressing, for example, the construct labelled extraversion and the construct labelled introversion depending on what situation they are in. It may be that some traits are generalisable across contexts and some are not. Or that some people are better than others at changing their trait expressions across contexts, and this in itself may be some individual difference which could be measured. This could answer questions like - is online personality the same as offline? Such a question would have implications for the results of studies predicting personality using social media (Kosinski et al., 2015). If behavioural expressions cannot be adequately described or predicted, using the personality constructs available to researchers, they may need to be re-evaluated. Goldberg (1990) Goldberg used 1,431 trait terms to define the Big Five, but now there are petabytes of descriptions of people's behaviour in big data to analyse and assess structures (using unsupervised machine learning algorithms).

In light of this potential for big data to capture psychological qualities, the following study investigates the possibility of measuring anxiety (usually assessed using self-report questionnaires) using the purchases of over the counter anti-anxiety medications in transaction data.

Anxiety

Anxiety and depression are the two most common mental health problems (Spitzer et al., 1994; Leon et al., 1995; Ormel et al., 1994; Olfson et al., 2000; Ansseau et al., 2004). Although they are highly comorbid (Moras et al., 1996), much of the focus has been on understanding and treating depression - with less efforts (in terms of research, public screening, and treatment) focused on anxiety (Kroenke et al., 2007). This is unfortunate as anxiety affects nearly 1 in 5 adults in the U.S. (Kessler et al., 2005) and can bring about devastating consequences, not just for the individuals affected but also for greater society. The symptoms of anxiety which can include excessive worry, abnormal autonomic arousal, sweating, palpitations, shaking, difficulty breathing, fear of losing control, an inability to relax, and sleeping difficulties (American Psychiatric Association, 2013) can substantially and persistently impact on daily functioning and work productivity. The economic cost in the U.S. alone is estimated at \$42 billion dollars a year, accounting for both direct and indirect costs (Greenberg et al., 1999). If anxiety can be diagnosed and treated earlier, the impact of the disorder can be minimised.

As well as diagnosed generalised anxiety disorders, many people experience particular symptoms of anxiety disorders, such as a difficulty sleeping or over worrying. These symptoms may be experienced on their own, or not be deemed severe or persistent enough to warrant the prescription of anti-anxiety medication. Additionally some people may actively seek to avoid prescribed medication as a treatment for anxiety due to stigma, fear of an over reliance on medication, or because of the potential side effects (Leonard, 2018). Instead, some people choose to purchase 'softer' or herbal remedies over the counter to help them

relax or sleep better. This consumer behaviour, can be measured in transaction data, and be harnessed as a way of studying anxiety symptoms in the general population.

Personality and Anxiety

Personality is receiving an increasing amount of attention in the health domain, as it has been shown to predict and influence various different health outcomes and behaviours. For example, personality has been shown to affect patient adherence to medication (Williams et al., 2011), longevity and the development of disease (Weston et al., 2015), and even the speed of physiological recovery after surgery (Mathews and Ridgeway, 1981).

In the mental health domain, certain traits and facets have been identified as comprising a ‘healthy’ well-adjusted personality type (Bleidorn et al., 2019). Conversely, trait N has been shown to negatively predict both mental and physical health (Bleidorn et al., 2019; Mathews and Ridgeway, 1981; Lahey, 2009). Poor regulation of emotion can affect physiological and mental stress arousal, which can lead to maladaptive health behaviours (Mathews and Ridgeway, 1981; Lahey, 2009). For example, Calkins et al. (2013) in a stepwise regression showed that N was a more important predictor of sleep disturbances than using the Anxiety Sensitivity Index (ASI) or the Dysfunctional Beliefs and Attitudes about Sleep Scale (DBAS-16) measures.

Clinical anxiety has also been robustly associated with high levels of N (Brown and Barlow, 2009; Kaplan et al., 2015; Kotov et al., 2010). Further, studies have found clinical anxiety to be negatively related to E (Kaplan et al., 2015; Kotov et al., 2010) and Conscientiousness (C) (Kotov et al., 2010). Facets of impulsivity have been associated with generalised anxiety disorder (GAD) (Pawluk and Koerner, 2013). In a study looking at anxiety disorder symptoms in non-clinical children, Muris et al. (2009) found that higher levels of N and lower levels of E were associated with higher levels of anxiety symptoms. Overall, it is less well known how personality traits affect anxiety symptoms out of clinical samples to those in the general population.

Anxiety in transaction data

Prescription or GP records are highly sensitive data, although they could be used to gain ground truth information on individuals with depression and anxiety, many precautions need to be taken to ensure records remain secure and anonymous, and even then, researchers having access to the data can make people feel uncomfortable. Over-the-counter transaction data, however, doesn’t hold as sensitive information as medical records, and the individual has given permission for their data to be used when signing up for the loyalty card, making it more suitable for research.

People who have severe or clinical anxiety or depression may be taking prescription drugs. These people may also exist in the Boots sample if they are purchasing over the counter anti-anxiety medication in addition to prescription drugs. In general, it is expected that people who buy over the counter anxiety medications are likely to be experiencing milder symptoms

of anxiety. Thus, it is proposed here that this data could be used to measure and understand more subtle anxiety symptoms in the general (non-clinical) population.

Predictions

If the purchasing of anxiety products has the potential to be used as a behavioural measure of anxiety symptoms in the general population, then it is expected that the behaviour is associated with the same psychological variables as self-reported anxiety symptoms in the literature. E and N are robust predictors of clinical and non-clinical anxiety in the literature. Therefore, it is predicted that N (positively) and E (negatively) will be associated with the purchasing of anti-anxiety products at Boots.

Although some people who are buying anxiety products in Boots may have a clinical disorder, it is likely that most customers buying anti-anxiety medication are doing so as an alternative not a supplement to prescription drugs. Therefore, the effect sizes of the associated traits with the purchasing behaviour are expected to be smaller than those found in a clinical setting. Given this, if the expected trait associations are observed, this supports the use of transaction data as a measure of anxiety symptoms in the general (non clinical) population.

In addition, this study investigates to what extent it can be predicted whether someone will buy anxiety relief not using all the pooled information gained from the survey questions provided by the participant. The survey asked participants about other psychological constructs (such as time preferences, frugality, self control, well-being, impulsivity), demographic information, and about their shopping motivations. Thus, this study also comprises an exploratory investigation into other psychological and demographic factors that significantly predict the purchasing of anxiety medication.

8.1.1 Method

Participants

Participants were from the sample of customers who's completed survey responses were matched to their loyalty card data, as described in Chapter 3, Section 3.2. Participants provided permission - through informed consent - for researchers to access four years of their loyalty card transaction data. The transaction data was pseudo-anonymised to link self-report questionnaires with the individual's loyalty cards number. No incentives were offered for taking part in the study. This study was approved by the School of Computer Science, University of Nottingham ethics committee.

From the 12,968 surveys that were returned and matched loyalty card transaction data, a sub-sample was extracted. To attain this sample, first the people who had purchased two or more (to eliminate the noise signal of one of 'experimental' purchases) of the 25 predefined anxiety products were identified. These anxiety items were defined based on previous research identifying them as being related to Seasonal Affective Disorder (SAD)

(Dzogang et al., 2017). The full list can be found in the Supplementary Materials (Section 10.5.1). Therefore, the anxiety purchasing sample consisted of 674 questionnaire respondents. These were matched at random to a sample of customers who had not bought anxiety items. This gave a balanced binary sample of 1,348 respondents from the UK.

Materials

The survey used in this study was the generalised survey described in Chapter 3, which was designed with the intention of investigating which psychological factors affect different purchasing behaviours. The survey comprised of questions on socio-demographic and psychological factors including time preferences (Reimers et al., 2009), self control (Tangney et al., 2004b), Behavioural Activation System (BAS), and the Ten Item Personality Inventory (TIPI) version of the Big Five (Gosling et al., 2003). Further items measured shopping motivations including an eight item frugality measure (Lastovicka et al., 1999), a shopping impulsivity scale (Rook and Fisher, 1995), and ten questions on shopping motivations. A link to a full copy of the survey sent to participants can be found in the Supplementary Materials (Section 10.1.1).

Analysis

The first analysis involved identifying whether any significant difference existed on the Big Five personality traits using multiple pair-wise t-test. The second analysis involved using all the survey questions input into a prediction model to test to what extent over the counter anxiety medication can be predicted on out-of-sample data, and what are the most important variables. For this second analysis a logistic regression model was chosen to ensure the results were highly interpretable.

Before being entered into the logistic regression model, categorical variables were coded as dummy variables (giving a total of 95 predictors). Because of income's ordinal nature, the lowest income group 'less than £25,000' was removed as the reference class. For the rest of the variables the most frequent class became the reference class: for loyalty card use this was 'every time', for shopping frequency 'once a fortnight', for the people they were living with this was 'partner', for marital status 'married', for occupation 'intermediate manager/professional/administrative', and for the highest qualification 'bachelors' degree, for loyalty card use it was 'Every time', for shopping frequency 'Once a Fortnight', and for region the most common response was to leave no response.

Variable Inflation Factors (VIFs) were checked, and indicated that happiness and satisfaction were highly correlated. These were subsequently combined into one well-being variable. Other variables which were correlated were 'Number of children' and 'Living with a partner and children'; and 'Marital status: Single' and 'Living with: Solo'. As the interest was in how factors in an individual's life may influence levels of anxiety, and as it cannot be assumed, for example, that all people who have children live with a partner and their children, these variables were left as they were for this analysis.

The second analysis utilised an inductive data-driven approach, but with the constraint that the final model would be a logistic regression, so the predictive model could be used to interpret the significant predictors of anti-anxiety purchases. First, 20% of the sample was excluded at random and set aside as the test set to evaluate the models on later. The remaining 80% of the dataset were used as training data. On this set the models would learn the patterns in the data (Bishop, 2006). To optimise the pattern recognition different meta parameters can be tuned. The optimal parameters for each model were found using a grid search. The grid search algorithm repetitively evaluated a range of different combinations of possible parameter values, fed to the algorithm in a grid format. After trying all possible combinations, the algorithm selects the parameters which produce the most accurate model (evaluated here by F1 score). The training and evaluation of the parameters were done using K-fold cross validation on the training data (80% sample), where $K=5$ (Bishop, 2006). K-fold cross validation (see Chapter 2, Section 2.1.4) was used to recycle data for training and testing in this validation phase.

Once the logistic regression model was optimised on the training data, its performance was tested on the unseen (20%) test data. The models' accuracy was evaluated using the F_1 score. Testing the models on data that is unseen by the algorithm ensures that the model generalises and isn't over-fitting to the sample. Despite deciding prior to analysis to use a simple and statistically interpretable model, post-hoc, non-linear models (decision trees and random forests) were fitted to the data to investigate whether the prediction accuracy was constrained by the linear nature of the logistic regression model. However, neither significantly improved prediction accuracy, suggesting the underlying structure of the data was linear.

Variable Importance

There are several ways in which variables can be ranked in importance from a logistic regression (Thompson, 2009). Here, standardised β coefficients were used, which assess the effect that the independent variable has on the dependent variable whilst all other variables in the model are held stable. The advantage of this method is that it doesn't penalise binary variables in comparison to continuous variables, and the significance of the variables can easily be assessed using p -values. Before extracting the standardised β coefficients, the logistic regression model was fitted to 100% of the data, keeping the same meta-parameters, to give the most information available whilst still protecting against over fitting.

8.1.2 Results

Firstly, t -tests were performed to see whether there were differences in the big five personality traits between those who bought anxiety medication and those that did not. Those who bought anxiety items were significantly less emotionally stable ($t(1346) = 5.04$, $p < 0.001$, $d = 0.27$), and significantly less extrovert ($t(1346) = 2.50$, $p = 0.013$, $d = 0.14$). None of the

other Big Five traits differed significantly between those who bought anxiety items and those who did not.

The optimised logistic regression model was then used to investigate the extent to which over the counter anxiety medication can be predicted using the self report measures from the questionnaire. The model was trained on 80% of the data, and tested on 20%. The test accuracy was 0.58, an 8% improvement on the baseline. This equated to a McFaddens R^2 of 0.10, or 10% of the total sample's variance explained. The significant predictors, β coefficients and 95% confidence intervals can be found in Table 8.2. A full table of results can be found in the Supplementary Materials (Section 10.5.2).

Post-hoc Analyses

Extraversion (E) and N were the only significant psychological constructs that predicted anxiety purchasing, surprisingly well-being was not. This may be because the purchasing of anxiety medications is more associated with daily happiness and not the satisfaction with life facet of Subjective Well-Being (SWB); or because well-being and trait N were correlated ($r = 0.39$), and thus N could be masking the effect of well-being. Post-hoc exploratory t -tests was performed to try to unravel these theories. There was a significant difference in well-being between those who bought anti-anxiety medication, and those that did not ($t(1346)=2.38$, $p<0.05$, $d=0.13$) suggesting that the effect of N could have masked the effect of well-being in the regression model. However when well-being was broken down into happiness and satisfaction with life facets, there was a significant difference in happiness between buyers and non buyer of anxiety medication ($t(1346)=2.68$, $p<0.01$, $d=0.15$) but no differences in satisfaction ($t(1346)=1.79$, $p=0.07$, $d=0.10$).

In the logistic regression, the only income group that was predictive of anxiety purchases was the group £75,000 to £99,999. This income group positively predicted purchasing anxiety relief. This high income predictive group was counter to intuitions that people with less money, have more stresses in life, and thus may buy more anti-anxiety medication/ sleep aids. To investigate the effect of income on anti-anxiety medication purchasing further, the ratio of those who purchased versus those who did not, were broken down by income group. These can be viewed in Figure 8.1. Comparative to people who did not purchase anti-anxiety medication, the £75,000-£99,999 income group had noticeably more people who did purchase the medication. However, with less people in these income groups, this effect could have been a consequence of the random matching of those who bought anxiety medication to those who did not - which wasn't matched on income. Therefore, the normalised anxiety sales were calculated to identify how much people spent in each income group relative to their total spend at Boots (as a percentage spend on anxiety items). These results can be found in Figure 8.2. The normalised anxiety sales show that the lowest income group (less than £25,000) spent a greater proportion of their spend at Boots on anti-anxiety medication than any other (0.25%) equating to an average of 4.7 items per person costing £26.57 over a period of 4 years. The trend then continues with the lower income groups spending a greater

proportion of their money on anxiety relief medications, with the exception of the highest income group.

Table 8.1: Sample descriptive statistics for the continuous predictor variables

	Mean	St. Dev.	Min	Max
Age (years)	51.88	14.63	18	87
No. of children	0.27	0.66	0	4
Openness to experience	4.82	1.20	1	7
Conscientiousness	5.89	1.05	1	7
Extraversion	3.91	1.57	1	7
Agreeableness	5.59	1.07	1	7
Emotional stability	4.54	1.49	1	7
Selfcontrol	3.35	0.64	1.23	5
Frugality	4.58	0.90	1	6
Shopping impulsivity	2.70	0.86	1.11	5.11
BAS-RR	3.18	0.47	1	4
BAS-FS	2.71	0.52	1	4
BAS-D	2.39	0.63	1	4
Subjective Well-being	7.02	1.98	0	10
I would rather stick to a brand...	4.54	1.57	1	7
I enjoy taking chances in buying...	3.94	1.54	1	7
If I like a brand I rarely switch...	4.67	1.52	1	7
I would not mind paying more...	5.42	1.23	1	7
I only buy products I trust...	5.31	1.38	1	7
Product Quality is important...	6.15	1.02	1	7
I am pleased about how I look...	4.26	1.56	1	7
When I feel good about my looks...	5.79	1.20	1	7
Whenever I see a mirror...	3.86	1.76	1	7
Environmental considerations...	4.54	1.48	1	7
I am concerned about climate...	5.13	1.54	1	7

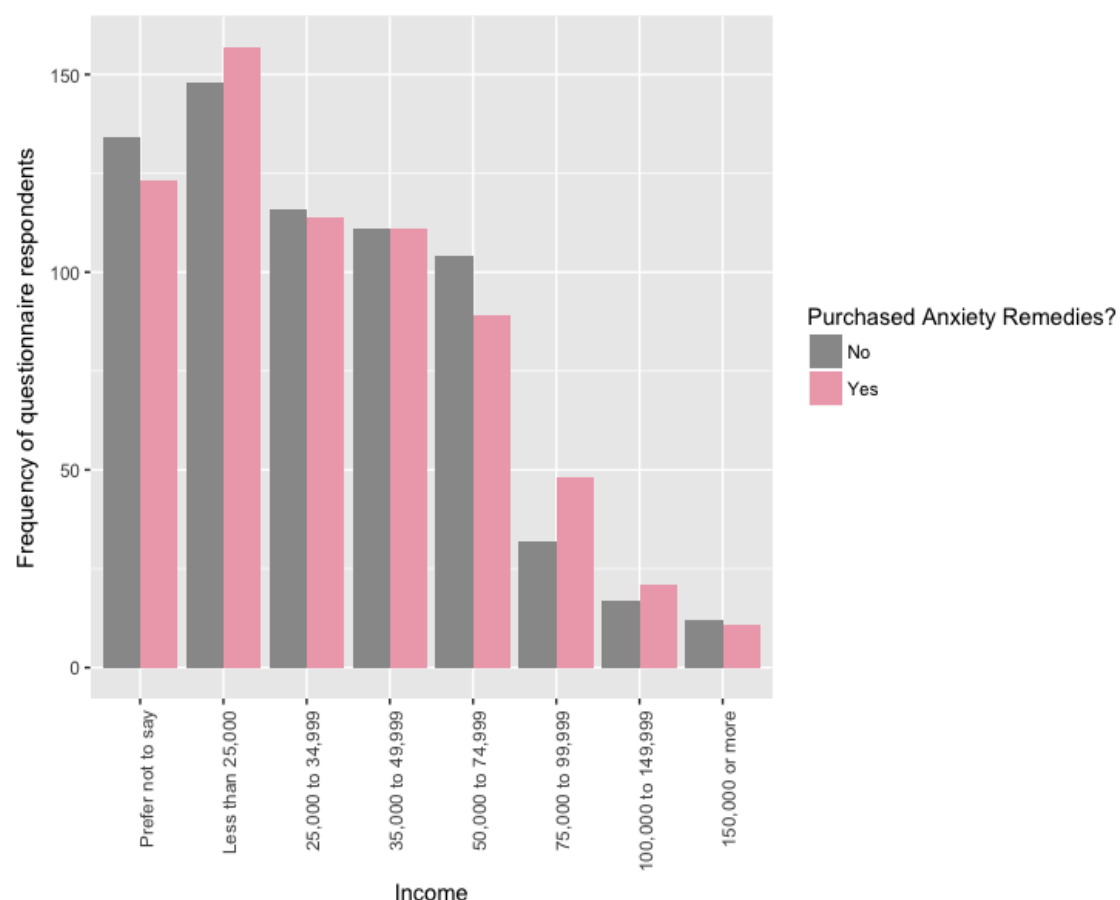


Figure 8.1: A graph to illustrate the number of survey respondents who did/did not purchase over the counter anti-anxiety products for each income group.

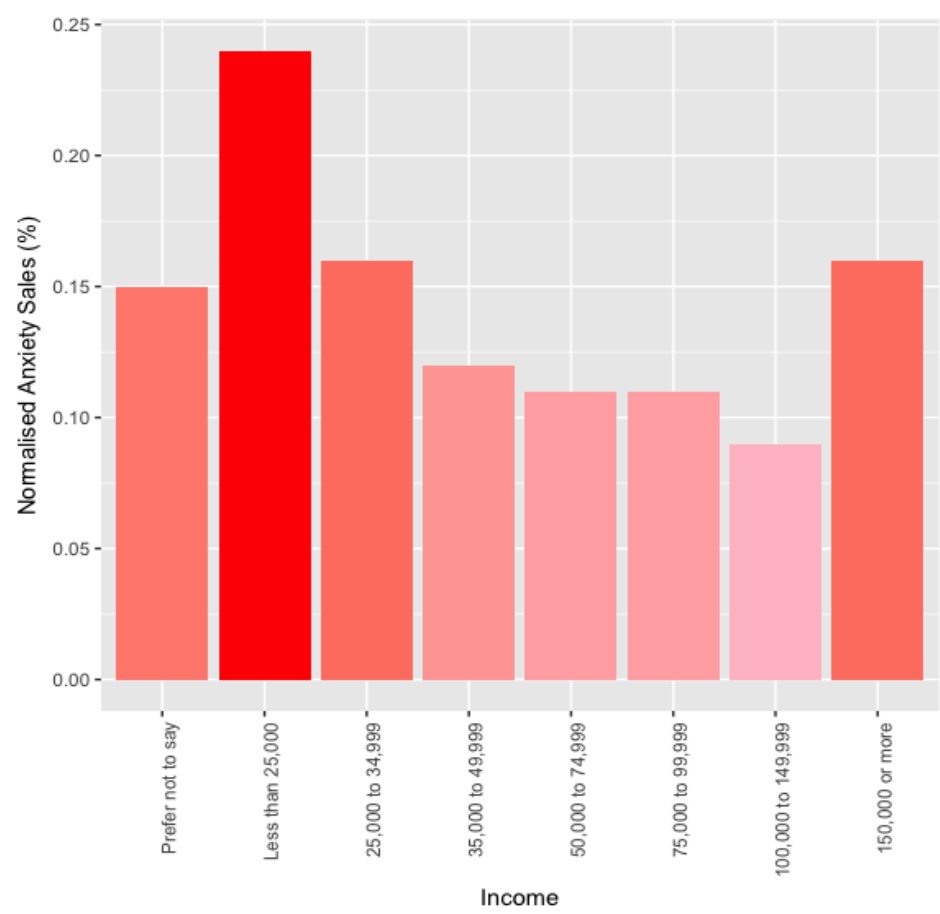


Figure 8.2: A graph to illustrate the percentage of sales in each income group that purchased over the counter anti-anxiety products.

Table 8.2: Logistic regression coefficients predicting the purchasing of over the counter anxiety medication

	<i>β coefficient (95% CI)</i>
Shopping freq.: Don't know	−2.31** (−4.62, −0.004)
Shopping freq.: Once every 4-6 months	−2.17** (−4.11, −0.23)
Shopping freq.: More than twice a week	1.07*** (0.39, 1.75)
Region: North East	−1.00** (−1.78, −0.21)
Gender: Male	−0.81*** (−1.28, −0.34)
Would rather £25 now to £75 in 3 months	−0.69*** (−1.18, −0.20)
Region: Northern Ireland	−0.69** (−1.35, −0.04)
Income: £75,000 to 99,999	0.53* (−0.10, 1.16)
Card usage: Most of the time	−0.52** (−0.93, −0.11)
Shopping freq.: Once a week	0.32** (0.001, 0.64)
Shopping freq.: Once a month	−0.31* (−0.64, 0.02)
Shopping impulsivity	0.25*** (0.06, 0.45)
Emotional stability	−0.20*** (−0.30, −0.10)
Environmental considerations...	0.12** (0.02, 0.23)
Extraversion	−0.12** (−0.21, −0.03)
I am concerned about climate change...	−0.10* (−0.20, 0.003)
Age (years)	0.01** (0.002, 0.03)
Observations	1,348
Log Likelihood	−837.06
Akaike Inf. Crit.	1,862.11

Note: *p<0.1; **p<0.05; ***p<0.01

8.1.3 Discussion

As predicted it was found that people who bought over the counter anxiety medication were lower in ES and E, replicating previous findings in the literature. This suggests that buying over the counter anti-anxiety medication is carried out by people of a similar personality disposition as those who report anxiety symptoms in questionnaires. This gives the purchasing behaviour some face validity as a behavioural measure of anxiety symptoms.

As expected, due to the purchasing of anti-anxiety medication measuring a prevalence of less severe anxiety symptoms than those experienced within a clinical population, both effect sizes were small. None of the other Big Five personality traits varied significantly between the two groups, as expected given the mixed findings in clinical populations (Kaplan et al., 2015). C was not a predictor of the anti-anxiety purchases, which had been found by a few studies in the literature to be negatively related to clinical anxiety (Kotov et al., 2010).

The logistic regression model, could predict 8% better than a mean baseline prediction, evidencing that the self report questionnaires carry some useful information about the differences between people who buy, and do not buy, anxiety medication. However, if this model was being applied purely on the merit of its prediction performance for novel data, for example to identify those who are going to buy the medication (and thus are displaying symptoms of anxiety), it would not produce particularly accurate results. If pure prediction accuracy was the applied goal, then incorporating information on what other product and brands people bought could significantly improve prediction. For example, exploratory analysis showed that the purchasing of anxiety medication was highly associated with buying lavender products and calming essential oils.

The strongest predictor of over the counter anxiety medication was shopping frequently at Boots. Interestingly, those who did not know their shopping frequency were the least likely to buy anxiety medication. This could be because they shop so infrequently they haven't kept track of their purchasing. Those who shopped more regularly than once a week were more likely to buy anxiety medication, and those who shopped less often than once a month were less likely to. All other factors ignored and assuming nothing about the item: if someone is present more frequently in a shop, there is a greater prior probability they will buy that item, than if they visit the store less frequently. Therefore, it is expected that an increased shopping frequency predicts an increased likelihood of buying anti-anxiety medications. However, this effect may be particularly strong if those who shop regularly, and thus consume more products, are more likely to perceive a product as being able to fix their problems (such as poor sleep or the inability to relax). They may perceive, for example, sleep aids to be a fix to sleep problems, the same as beauty products fix things with their body they are not happy with. Thus, this effect may be due to an underlying reliance, or greater inclination, to use consumption as a solution to things they are dissatisfied with. This theory aligns with the finding that shopping impulsivity positively predicts anti-anxiety purchases. This is something that should be studied further if the purchasing of over the

counter anxiety medication is going to be developed as a behavioural measure of anxiety, as the implications are that it may be confounded by consumerism and shopping impulsivity.

Answering affirmative to the question ‘would you rather have £25 now or £75 in 3 months?’ significantly negatively predicted the purchasing of anti-anxiety medication. Therefore, people who are more present-focused are less likely to buy anxiety remedies. It is speculated that, assuming there are a greater number of potential things to worry about in the future compared with the present, someone who is present-focused has less potential to unnecessarily worry and develop symptoms of anxiety, compared with someone who is future focused.

The logistic regression model showed that males were less likely to buy anxiety medication than females. This effect replicates previous findings showing that females are higher in trait N (Lynn and Martin, 1997), and more prone to anxiety disorders (Lewinsohn et al., 1998; Angst and Dobler-Mikola, 1985; Bruce et al., 2005; Regier et al., 1990; Kessler et al., 1994).

The post-hoc analyses showed that there was a significant difference in happiness, but not satisfaction with life, between those who bought anti-anxiety products, and those that did not. Thus, although these facets are highly correlated, they relate differently to anxiety symptoms and the purchasing of anxiety relief. This has important implications for well-being and anxiety research as it implies that people can evaluate their satisfaction with life not taking into account their emotional well-being. For example someone may have a highly stressful job, causing symptoms of anxiety, but because it is well paid or prestigious may report being satisfied with their life.

The logistic regression found that the income group £75,000 to £149,999 was predictive of buying anxiety items or not (see Figure 8.1). However, when the effect of income on anxiety purchases were normalised by the customer’s total spend, the lowest income group (less than £25,000) spent a greater proportion of their spend at Boots on anti-anxiety medication than any other (0.25%) (see Figure 8.2). This equates to an average of 4.7 items per person, an average of £26.57 over a period of 4 years. These findings suggest that people with a low income have greater anxiety symptoms, and that in general a higher income can guard against anxiety symptoms, until income reaches a certain threshold, at which point presumably the type of job required to yield such an income is stressful and anxiety provoking. Unfortunately, with the broad income categories, there was not enough fine grained information to ascertain what income amount increases the purchasing of anxiety items.

In the literature, facets of impulsivity have been associated with generalised anxiety disorder (GAD) (Pawluk and Koerner, 2013), however this study found that trait impulsivity was not a significant predictor of anti-anxiety medication. Therefore, the behaviour of buying anxiety medication to treat anxiety symptoms is likely affected by different mechanism than GAD. Although trait impulsivity was not a significant predictor, shopping impulsivity was. This suggests that some form impulsive nature drives the purchasing of anxiety items, but it is constrained to a consumer context. This finding also supports the notion that the purchasing of anxiety items is a consumerist solution to the problem of emerging anxiety

symptoms. Thus, those who express their impulsiveness through consumption, also express their anxiety via consumption. The positive relationship with shopping impulsivity may also suggest that the purchasing of anxiety items may not be a planned behaviour. Further analyses could investigate this by analysing the entropy (or randomness) of the time intervals between purchases (to see if they are bought as part of repetitive cycles or not); as well as the type of baskets these medications fall into - whether they are more frequently bought as part of a regular 'stock up' shopping mission, or if they are purchased with random items or items known to be impulsive buys.

The extent to which someone's environmental considerations affect the products they purchase is a positive predictor of purchasing anxiety medication, yet general concern over climate change is a negative predictor. As in Chapter 6 when being used to predict regular bag purchasing, it shows that these questions despite both being about environmental concerns tap into different, potentially orthogonal, underlying psychological processes/motivations. In this instance many of the anxiety products are considered herbal remedies, and thus people who buy products based on their environmental impact, may be more inclined to choose these sorts of herbal sleep aids and anxiety relief. Further research is required to understand the relationship between environmental attitudes and anxiety symptoms in the general population.

The regions of Northern Ireland and the North East were found to be negatively related to the purchasing of over the counter anxiety medications. This indicates that there may be some regional differences either in the prevalence of anxiety symptoms in the general population, or the way in which people deal with them (i.e. through consumption or not). However in some regions there were much smaller samples, and thus the generalisability of these findings is questionable. An analysis of the potential geographic variances in anxiety purchases are investigated further in Section 8.2.

Why Predicting Behaviour in Big Data is Challenging

That there were some differences in the psychological constructs predicting self-reported anxiety and the purchasing of anti-anxiety medicine, could suggest that we shouldn't be looking for a behavioural measure of anxiety to *replace* self-reported anxiety levels, but instead to *supplement* it. For example, the behavioural measure may be tapping into something qualitatively different. Low C individuals are more likely to have anxiety (Kotov et al., 2010), measured using self reports. Whilst the people buying anti-anxiety products were not low C, perhaps due to the behavioural measure requiring a proactive action (buying something to help the symptoms) may entail that it measures anxiety for a particular type of person; or a qualitatively variant type of anxiety. For example, if there were differences found between the individuals who have anxiety but seek help, and those who don't (or seek different types of help), this could have important implications in the mental health domain.

This study found smaller relationships between personality traits and anxiety expressed as a behaviour than has been found for self-reported anxiety symptoms. However, for anxiety

purchasing to be established as a behavioural measure of anxiety (or for any behavioural measure of a psychological phenomena to be established) the behaviour has to be predictable - illustrating there is some underlying psychological mechanism, and that it is not merely noise or randomness. The problem with this is that human behaviour in the real world is fundamentally hard to predict. Not only are researchers limited by not having complete data on individuals, or ground truth explanations attached to behaviours, the properties of human behaviour itself (that it is complex and influenced by a multitude of different variables) makes it hard to predict using psychological theories.

The reason's real world behaviour is hard to predict align with the limitations of psychology as an empirical science in the classical sense, as laid out in Smedslund (2016). Firstly, there is the problem of irreversibly when studying psychological processes in people. Every experience changes a person, and this cannot be completely undone due to conscious human memory and synaptic adaptation. Thus, this makes it hard for a single participant's responses to be tested in two different conditions, as one has to come before another and the first condition changes the person irreversibly. This makes the interpretation of results ambiguous as to whether any change was resultant from the person or from the conditions.

The experiences and the memories that make humans human, also make them incredibly hard to predict. Experiences allow people to be affected by an infinite amount of previous thoughts, memories, and emotions, all of which are invisible and difficult for researchers to measure and quantify. Thus, predicting individual behaviours is difficult, and highlights why out-of-sample testing of models is so key for a true evaluation of a models performance in the real world. For example, a prediction model may predict well for a response on one day, but by tomorrow the people who the model is predicting will have had new experiences which could influence their responses on the next day. This highlights how, when models are simply fitted to one set of data in time, the model is not being evaluated for empirical prediction accuracy.

Experiences change people, and the experiences people have can be infinitely different. This means that scientists are studying individuals in an infinitely different number of contexts. Past experiences are important predictors of future behaviour. And yet many of these past experiences will be unknown by the researchers. This makes psychology immeasurably more complex than traditional natural sciences. In chemistry, for example, for two molecules to react, all variables can be accurately measured and controlled. Parts of the same compound or chemical, unchanged by time, can be used again and again under the same conditions to replicate the effects of a reaction.

This problem of infinitely different contexts, not only makes real world behaviour hard to predict, it also makes it hard to interpret. For example, humans interpret ambiguous movements or gestures based on the context and their own experiences. If someone looks towards the door, taken in isolation this could mean an innumerable amount of things. However, if one knows the previous experiences and the context, one can start to interpret the gesture. However in big data analysis the researcher doesn't know the context or experience

of a person buying a particular product, for example, and thus can only make assumptions about the buying behaviour based on their own experiences and logic. Yet, on an aggregate level, any intuition on why people may be buying a product will likely be drowned out in a sea of different customer's individual experiences and contexts that are unknown to the researcher.

It is easier to predict human behaviour over aggregate groups, analysing behaviour as averages. Psychology, however, is about understanding individuals (Lamiell, 2015). But valid inferences cannot always be made from averages to individuals (Wagner, 1982). In a laboratory set up, where researchers study stimulus responses, the extraneous controls applied to the situation can make the prediction of behaviour easier. However, in doing so, human behaviour is being studied in isolation without any interaction between persons. Many will agree with the notion that "the most important parts of human life are interactions with other persons" (Smedslund, 2016). For example, how does a researcher suppose to understand aggression, trust, and jealousy - key causes of altercations, fights, and even deaths in human communities - by isolating them and studying stimulus responses in a controlled environment. Researchers, by removing their subjects from a social environment to study behaviour, impair their ability to predict and understand real world behaviour. By measuring real world behaviour in big data, sociability is encompassed within the model - yet prediction accuracy likely falls as a consequence. Not only now are we trying to model an infinite number of peoples experiences and how they affect people's behaviour, we are also trying to model the interactions between them.

Summary and Implications

The aim of the study reported in this chapter was to investigate the possibility of over the counter anti-anxiety medication being a behavioural measure of anxiety in transaction data. Like much research in big data, there were no exact ground truth labels for self-reported anxiety. Therefore, the behaviour was first tested for face validity to check whether it related to the personality traits that would be expected from the literature (using self-report measures). It was found that purchasing anti-anxiety items was significantly negatively related to extraversion and emotional stability, as predicted from the literature on self-reported anxiety. Interestingly however, the behaviour was not related to conscientiousness.

For a behavioural measure to be established there needs to be a strong signal in the data, so it is possible that the behaviour could be driven by underlying psychological mechanisms rather than randomness or non-psychological predictors. This was tested by predicting the behaviour using all the self report information available. The prediction improvement from the baseline was small, yet it was there nonetheless. To better understand the behaviour, the most important predictors were analysed, and the model interpreted. Using the logistic regression model, emotional stability and extraversion negatively predicted anxiety medication purchases, when gender, income, age, shopping frequency and other psychological variables were controlled for. Interestingly shopping impulsivity also predicted the behaviour, sug-

gesting that these were not planned purchases, and may be related to consumerism. A more in-depth analysis of the purchasing behaviour and income showed that people in the lowest income group bought more anti-anxiety medication, despite having less money, signifying that having less money is related to anxiety. Interestingly, the next income group which bought the most anxiety products was the highest income group (£150,000 or more). This could be indicative of high powered jobs causing anxiety, and why satisfaction with life may not have been predictive (as those who have successful careers, which cause greater anxiety, may still rate their satisfaction with life as good).

This study discusses the possibility of establishing a behavioural measure of anxiety. The results show that with more behavioural data, and more complex algorithms, there is the potential for the purchasing of anxiety items to be developed as a behavioural indicator of anxiety, with certain shopping characteristics, such as consumerism, taken into account. For example, it is important to note that not all people who display anxiety symptoms will respond by purchasing sleep aids from Boots. However, once behavioural patterns characterised in the data of those who can be inferred to be high anxiety individuals, these behavioural patterns can be used to predict others who are at high risk (who may not have brought the products, but are expressing the same characteristic behaviours in their shopping patterns).

To conclude, studying individual behaviour using aggregated data of people with unknown experiences, who are ever changing and can interact with other people and contexts in an infinite number of ways, makes predicting behaviour in the real world a challenge. There are the additional caveats of not having a complete picture of individual's shopping patterns (only at one retailer), and that it is not known if people are buying the items for themselves or not. Given these, to even find a small predictive signal in the transaction data using psychological constructs is promising. Thus, constructing behavioural indicators of psychological states and traits in big data, yet still very much in its prime, is a fruitful research area for big data psychology.

8.2 Geospatial Analysis in Big Data

One of the advantages of big data, proposed in Chapter 1, is that data can be used to study behaviour over a greater spread of different geographic regions. Much of big data, is geo-tagged, meaning it is attached to a particular location. This may be the location someone was when they posted a tweet, the route of their run traced using a sports watch/Fitbit, or the location of a store in which they purchased an item. In the previous section, living in certain regions such as the North East, and Northern Ireland negatively predicted the purchasing of anti-anxiety medication in a regression model. This may reflect sample or store differences, or perhaps some people in some parts of the country are genuinely less prone to anxiety? This could be because of environmental factors, or due to differences in peoples personality, making some have a lower predisposition to have anxiety symptoms.

Research has shown that personality traits can in fact cluster geographically (Rentfrow et al., 2013; Krug and Kulhavy, 1973; Plaut et al., 2002) and these clusters can be related to health factors (Rentfrow et al., 2015). Such findings have been found by linking personality surveys to participants' postcodes. But what about when participant's postcode data is not available? Can people's geo-tagged digital data be used as a proxy of where they reside? This section explores the utility of aggregating personality data by region, using the location of the Boots store people shop in, and the region those Boots stores are located in. Such a methodology allows for the investigation of whether Boots customers' psychological attributes (measured using the survey) cluster spatially in the UK. In addition, it can be studied whether spatial distributions of personality traits are related to the spatial distribution of behavioural data, such as over the counter anti-anxiety medication. For example, does a region where people are high in neuroticism have higher over-the counter anxiety sales?

A further advantage of geo-tagged data is that the spatial element can be used to join one dataset to another. Many datasets can not be joined on an individual level, due to both logistical and privacy reasons. Yet the relationship between two datasets can still be studied if they are aggregated and linked by time (to study temporal trends) or by location (using coordinates). Aggregated data is also much more likely to be open source, as there are no privacy issues when individuals cannot be identified (when aggregated at a high enough level). Thus data with a spatial component provides an exciting new way to study human behaviour by combining multiple data sources together.

This section first studies whether personality from Boots customers, aggregated by region, clusters spatially. It then studies the relationship between anxiety medication sales and customer personality traits at a store level. And finally, the section finishes with a demonstration of how the Boots data can be joined to other open source datasets by geographic location, in order to gain a greater understanding of what exactly it is the spatially aggregated Boots data is capturing.

8.3 Do Boots Customers' Personality Traits Cluster Spatially?

This section, uses a series of analyses to investigate whether spatial clustering of Boots customers exist in the UK. Often address information is not stored in research databases due to privacy issues. Therefore, if spatial effects can still be studied using the region the data activity occurred/was collected in as a proxy, this could prove valuable for researchers interested in conducting geo-spatial analyses. Thus, the following analyses investigate whether customers traits cluster using the location of the Boots stores in which customer's are shopping in, as a proxy of whether they live, to then aggregate customer's personality traits geographically by district.

8.3.1 Introduction

The notion that personality might be geographically clustered comes from three different research areas. The first, is about *social* influence - that cultures, daily routines, and lifestyles affect an individuals psychology (Hofstede and McCrae, 2004). The second, is that ecological features in an individuals physical environment affect people and how they think and act. For example, dangerous or unhealthy environments can motivate people to be more risk-averse (Murray and Schaller, 2014; Schaller, 2006), or peoples values are altered by climatic and economic (Van de Vliert, 2013). The third area of research covers migration patterns, particularly selective migration, the notion that people move to places which allow them to express their personality to the fullest and satisfy their psychological needs. For example, an extrovert may move to a big city, as there are more exciting things to do and see there. Similarly, a liberal individual who is high in Openness to New Experiences (O) may want to move to a place where there are more liberal practices, such as the legalisation of cannabis, for example. Furthermore, some personalities have been found to be more likely to migrate than others. Jokela (2009) showed that people who are creative (high O) and sociable (high E) are more likely to migrate than those low on the trait dimensions. And agreeable individuals are more likely to stay in their hometowns (Boneva et al., 1998). In summary, there is different evidence from different areas to suggest that geographic clustering of the Big Five may exist.

Studies in the US have indicated that robust geographical patterns of personality traits do exist (Rentfrow et al., 2013, 2008; Krug and Kulhavy, 1973; Plaut et al., 2002). However, the US is much more vast, allowing much greater opportunity for geographic variation than in the UK. In 2015, Rentfrow et al. exploited a large scale personality survey by the BBC to investigate whether personality clusters geographically in the UK. They found that indeed there were spatial clustering on each of the 5 trait dimensions. Further, they found that geographic trait values correlated with geographic health, education, and political differences when controlling for age and gender. Particularly, they found that E was associated with high education, income, and longer life expectancy's; Agreeableness (A) was associated with low-income residents, less university-educated residents, and low rates of violent crime; C was associated with votes for Conservative over Labour candidates, greater longevity, less violent crime, and reduced deaths from cancer and heart disease; N was associated with higher violent-crime rates, long-term health problems, and deaths from stroke, cancer, and heart disease; whilst O was positively related to votes for Liberal Democrats, university education, high income, same-sex couples, and rates of violent crime.

Geographic clustering of personality traits have also been investigated at a more local level. (Jokela et al., 2015) examined how both personality and life satisfaction are geographically within the city of London. They found that the strength of association between personality traits and life satisfaction depends on neighbourhood features, and that people's well-being depends, in part, on the interaction between the individual's personality, and particular characteristics of where they live.

8.3.2 Analytical method

The individual level information in both the Boots transaction data and the survey data is not attached to post code information. Therefore both personality information and sales data needed to be aggregated over the same spatial area. In the transaction data, the store number is attached to longitude and latitude. Thus, the sales from a given store can be pin-pointed to a particular spatial location on a map (see Figure 8.3 for the distribution of Boots stores in the UK). The personality information is harder to attach a spatial location to. Seeing as there is no accurate data on where the individuals live, this is approximated by using the Boots stores people shop at. There are several methods in which personality information could have been aggregated to store level, of which two were trialled. Both methods involved calculating an average personality for each store's customer base, and tying a particular personality to each store. The two methods are outlined in more detail below. Using the same methods, local authority district (LAD) trait values could also be calculated, aggregating data at a higher level using the stores which occur within districts.

Method 1 used the personality traits of people to calculate the store personality only when a store was deemed as an individual's 'home store'. Under this condition, only personality information contributes to the store's personality when it was the individual's most regularly visited store. This, may be the nearest store to where someone lives, or it maybe the store next to where they work. Either way, the store will be expressing a personality of the store's most frequently visiting clientele eliminating noise from individuals randomly visiting stores out of their routine. Once store personalities were attained, district personality's were calculated by averaging the store personalities in the given district.

Method 2 calculated the mean store personality using every shopper's personality who had visited the store over the 4 year period of data available. This method will allow the stores to express an overall personality of the store's clientele. Each person who shopped at the store, their personality contributed just the once, irrespective of how often they shopped there. District personality was calculated by simply aggregating over district, rather than store.

Spatial Clustering

Several analyses were carried out to evaluate the extent of geo-spatial clustering in the data, and where the clusters were occurring. Both methods of calculating the stores customer's personality were trialled. Method two produced more reliable estimates (with more data contributing to each store personality calculation) and thus only *Method 2* is reported here for the sake of space. To further maximise the amount of data, and to ensure there was enough data in each geographical location, local authority district (LAD) was used as the level of aggregation (data on the 380 district boundaries are available online at data.gov.uk). Thus the following spatial auto correlation and hot-spot analysis was conducted using *Method 2* for calculating district personality of everyone who had shopped in a store in that district.

This analysis involved three steps:

1) First, the degree of clustering was evaluated using spatial auto correlation analysis. This was done first on the global level, using Global Moran’s I, to capture whether neighbouring and near districts are more similar than districts further away.

2) Second, on the local level to identify any significant local clusters or spatial outliers, using the local indicator of spatial association (LISA) statistic on the Local Moran’s I.

3) Lastly, a ‘hot-spot analysis’ using Getis-Ord G_i^* estimate was used to identify significant hot-spots of the different traits to include the most information possible.

These analyses are outlined in more detail, and the equations given below.

Spatial Auto-correlations

Global spatial auto correlation, measured by Moran’s I, captures the extent of overall clustering that exists in a dataset. Global Moran’s I is constrained to a scale of -1 to 1, 0 indicates no spatial correlations, where -1 indicates negative correlations (i.e. like a chequerboard) and 1 indicates spatial clustering. The equation for Moran’s global I can be found in Equation 8.1 where N is the number of spatial observations indexed by i and j , X is the variable of interest; $\bar{X}$ is the mean of X , w_{ij} is the matrix of spatial weights (with zeros on the diagonal) and W is the sum of all w_{ij} . Moran’s Global I was used to measure the extent of clustering over the whole of the UK, for each of the Big Five separately.

The Local Moran statistic, suggested by (Anselin, 1995), is a way to identifying local clusters and spatial outliers. The formula for Local Moran’s I (see Equation 8.2) is similar to the global form, except that the permutation is carried out for each observation in turn. LISA statistics can be calculated using the Local Moran’s I, producing a pseudo p -value for each local auto-correlation, which can be used to assess significance. All local statistics were calculated using 999 permutations.

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2} \quad (8.1)$$

$$I_i = N(X_i - \bar{X}) \frac{\sum_j w_{ij} (X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2} \quad (8.2)$$

Hot Spot analysis

To identify areas or ‘hot-spots’ that were statistically high or low on the Big Five, a Getis-Ord analysis (Ord and Getis, 1995) was performed. The Getis-Ord statistic is derived from a point pattern analysis logic. It is calculated as the ratio of the weighted average of the values in the neighbouring locations, to the sum of all values. The statistic is computed using equation 8.3 where the denominator is constant across all observations (as the total sum of all values in the dataset). The G_i^* statistic returned for each feature is a z-score. For statistically significant *positive* z-scores, the z-score will be large, and the more intense

the clustering of high values (a ‘hot spot’). For statistically significant *negative* z-scores, the smaller the z-score is, and the more intense the clustering of low values (cold spot). Hot spot analysis’s was conducted in this way for each of the Big Five traits.

$$G_i = \frac{\sum_j w_{ij} x_j}{\sum_j x_j} \quad (8.3)$$

8.3.3 Results

Moran’s Global I was calculated to investigate the extent of overall clustering of each of the Big Five across the UK. For trait openness, Moran’s global I was 0.03, for conscientiousness 0.01, extraversion 0.02, for agreeableness 0.04, and for neuroticism 0.03. These results indicate that there is a very small amount of similarity clustering over the Big Five, and more so for traits A, E and N.

The Local Moran’s I was then calculated to identify local clusters and spatial outliers. The significant local indicator of spatial association (LISA) auto correlations are presented in figures 8.4a to 8.8b and indicating the type (i.e. similarity/dissimilarity) and significance level.

Finally, the Getis-Ord statistic was used to identify statistically significant hot and cold spots in the UK for the different Big Five personality traits. These can be found in Figures 8.9a to 8.13b. The graphs show that various statistically significant hot spots occurred for each of the Big Five. For example, the northern points of Scotland and the North West of Northern Ireland were significantly low in Openness; whilst Southern Eastern parts of Northern Ireland and Eastern English districts in Lincolnshire and Norfolk are significantly higher. In addition, there is a significant cluster of low C in Western England (near Hertfordshire and Worcester), and a high C cluster North East of London. In the South East of Scotland, and the South West regions of Northern Ireland, are significantly lower in extraversion, whilst Lincolnshire in the East Midlands is significantly higher. In Western Scotland, levels of agreeableness are significantly high, and in West regions of Northern Ireland and clusters in London were significantly lower. Western England and Ireland were significantly lower in emotional stability, whilst North West Devon, North East of London, regions in the North and East Midlands, and the regions surrounding Edinburgh were significantly higher. However, these effects were relatively small, and when the bonferroni corrected p -value was taken ($p < 0.00013$), there were no significant hot-spots for any of the Big Five traits.

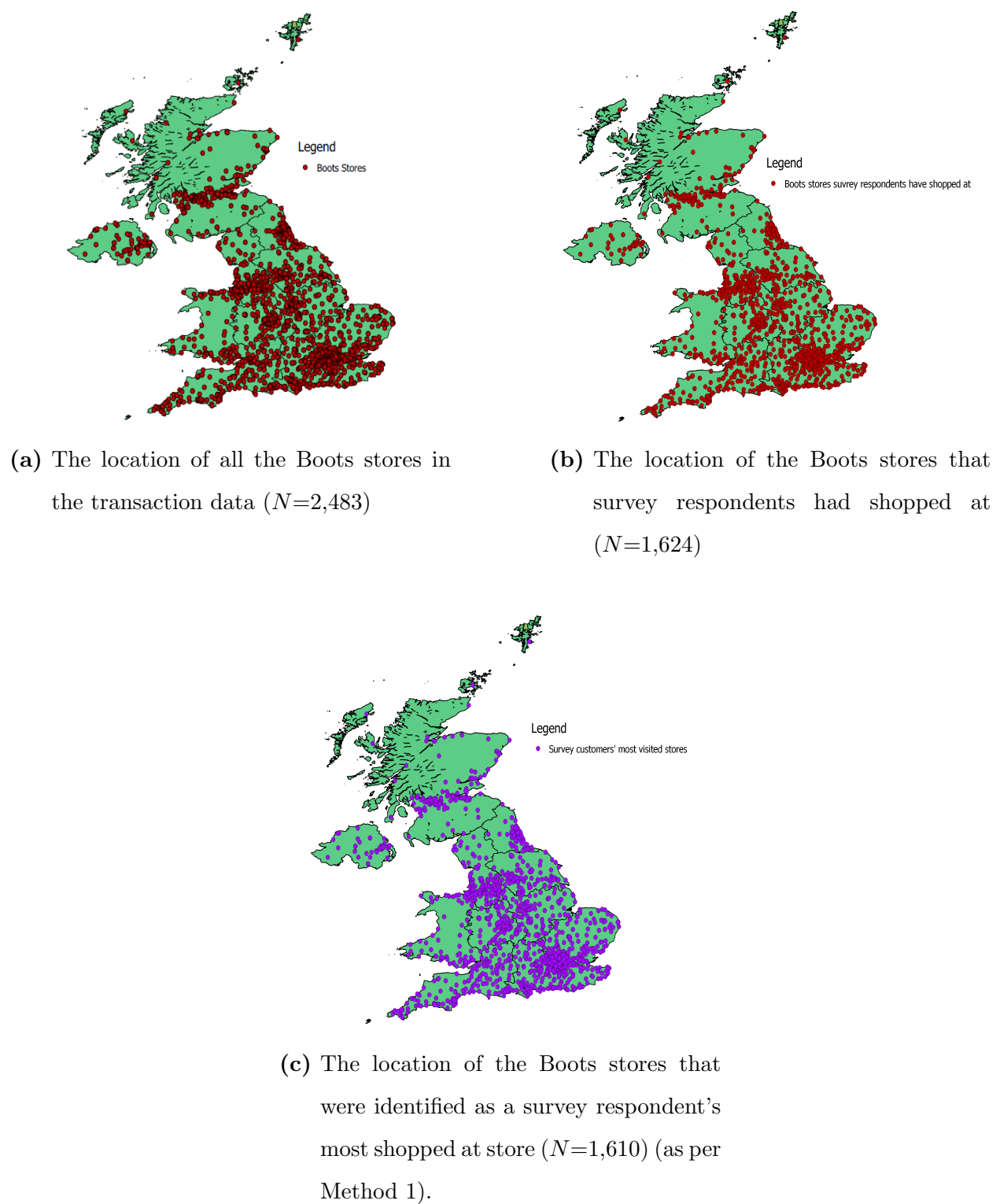
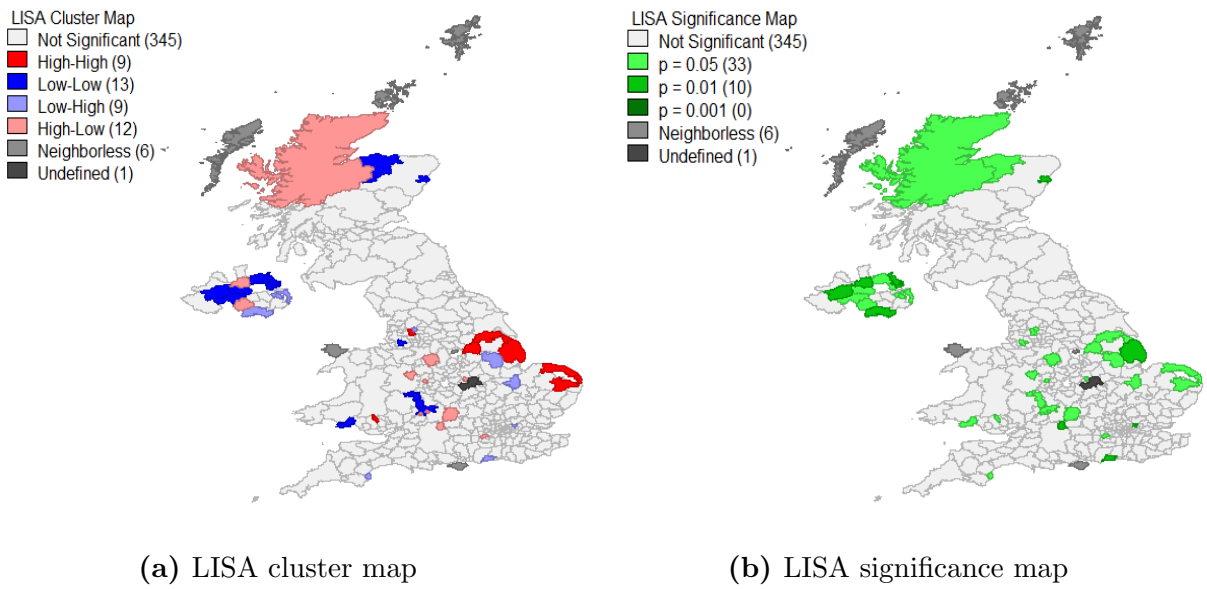


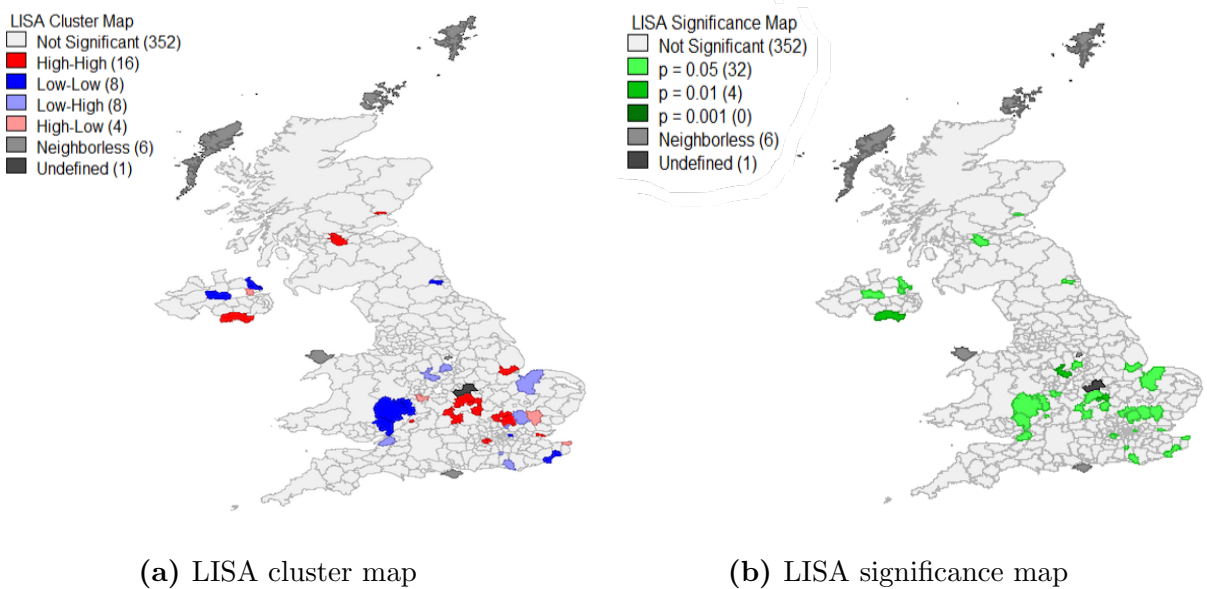
Figure 8.3: Maps to illustrate the distribution of Boots stores in the different samples of data used



(a) LISA cluster map

(b) LISA significance map

Figure 8.4: Openness



(a) LISA cluster map

(b) LISA significance map

Figure 8.5: Conscientiousness

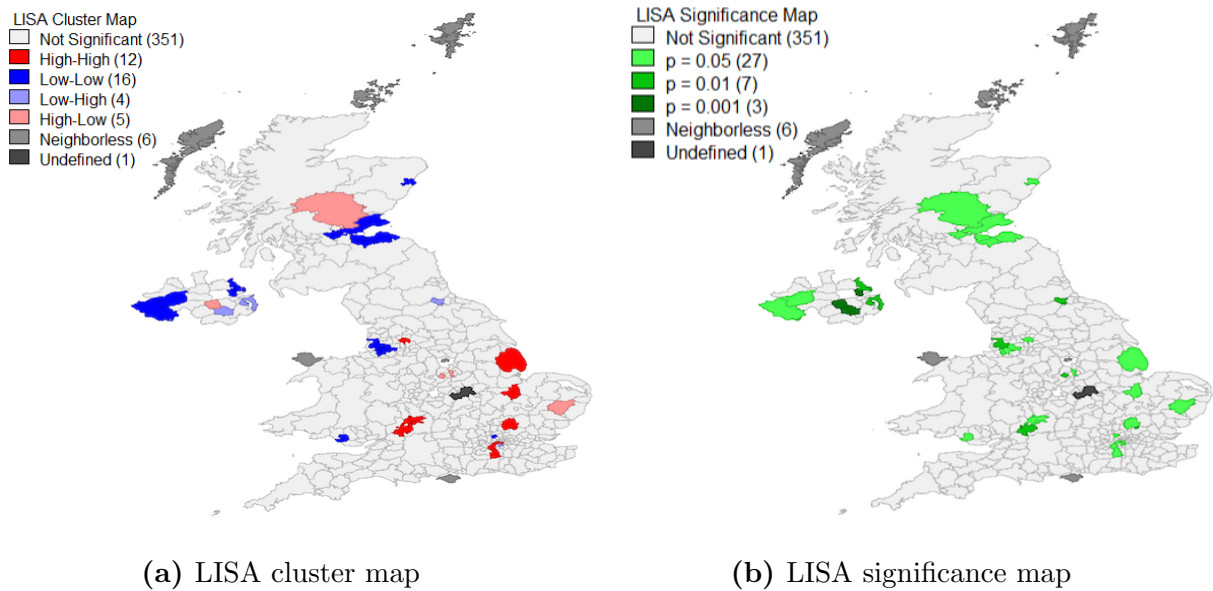


Figure 8.6: Extraversion

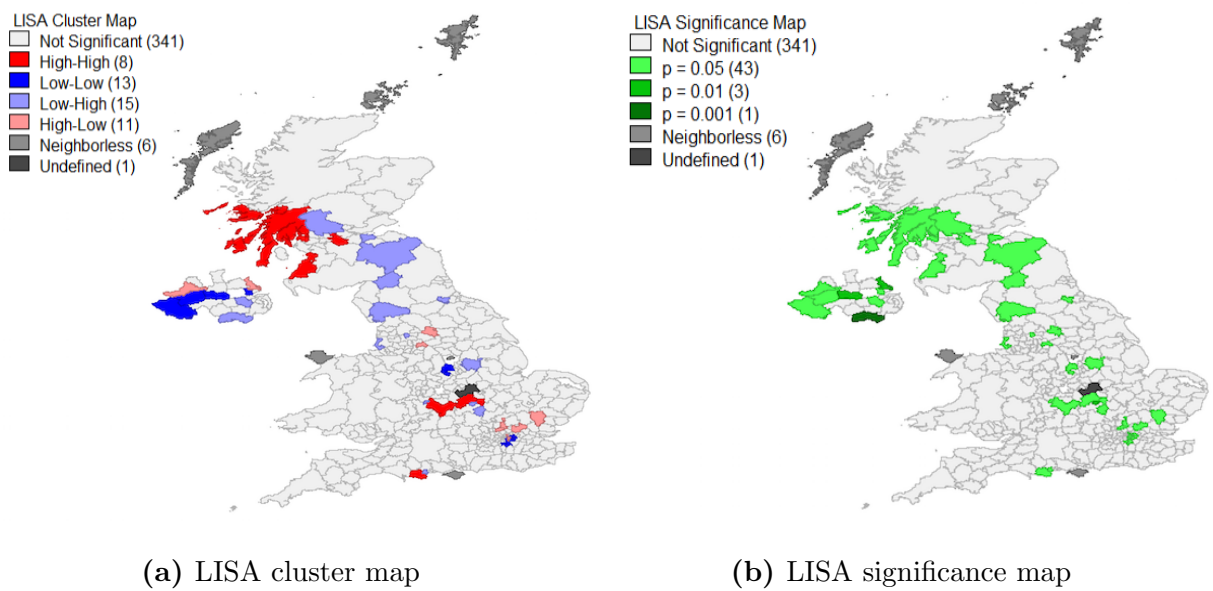


Figure 8.7: Agreeableness

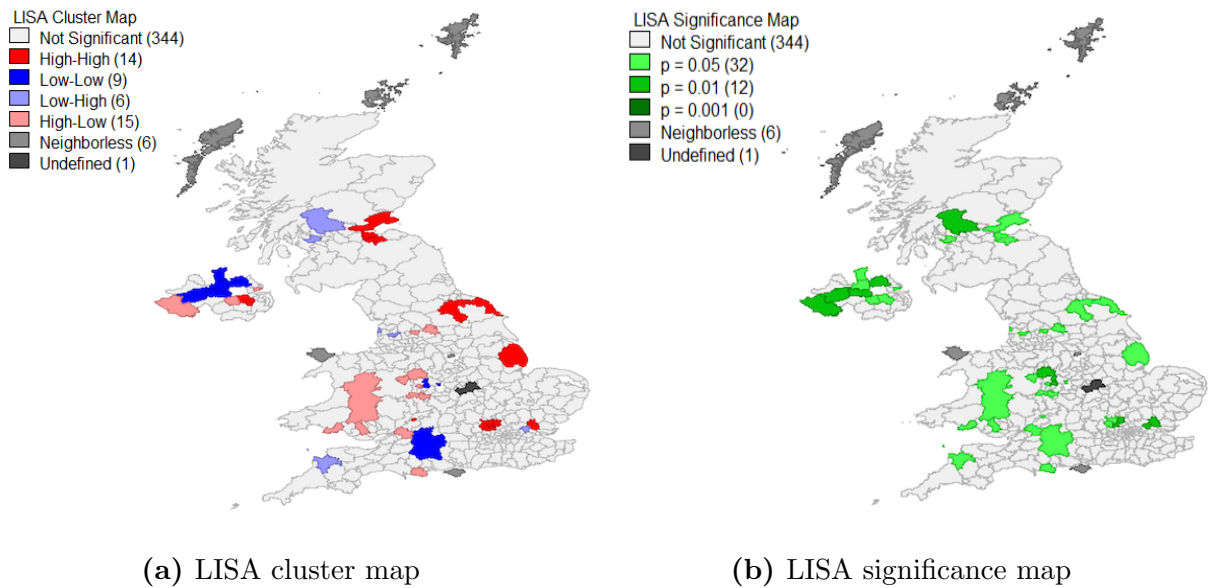


Figure 8.8: Emotional Stability

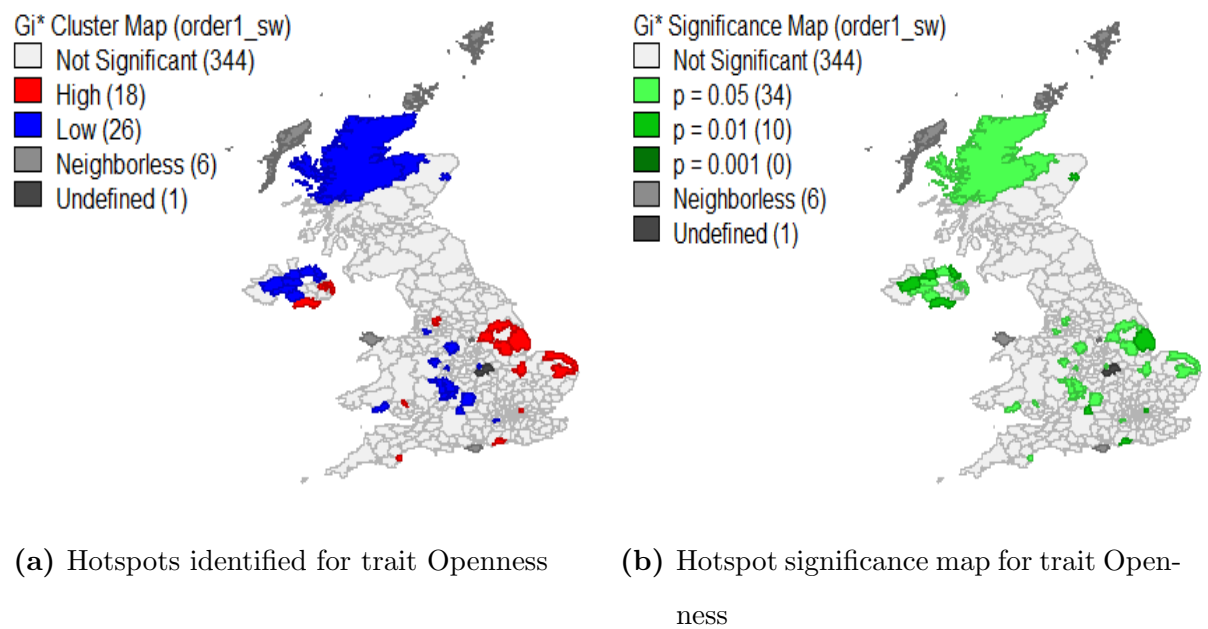


Figure 8.9: Openness

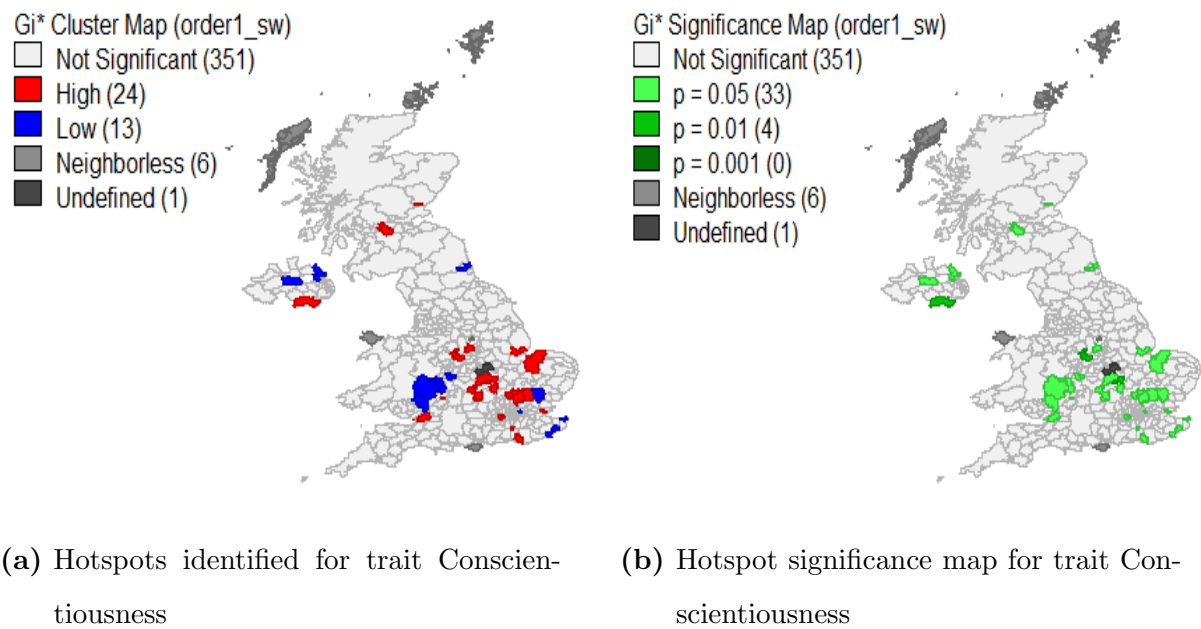


Figure 8.10: Conscientiousness

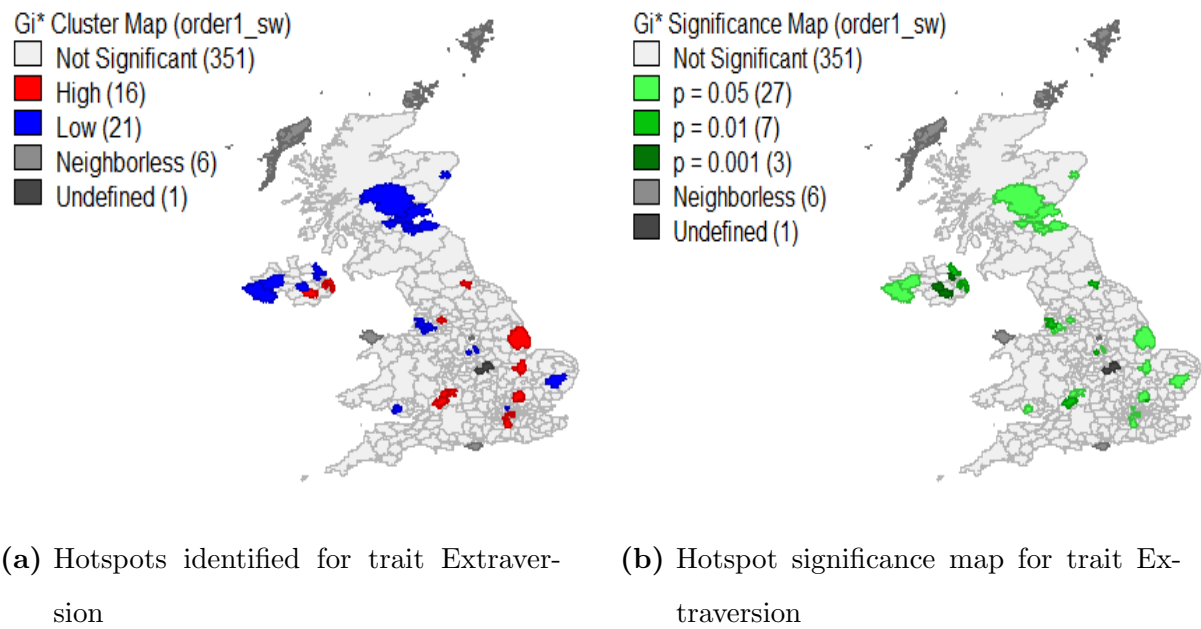


Figure 8.11: Extraversion

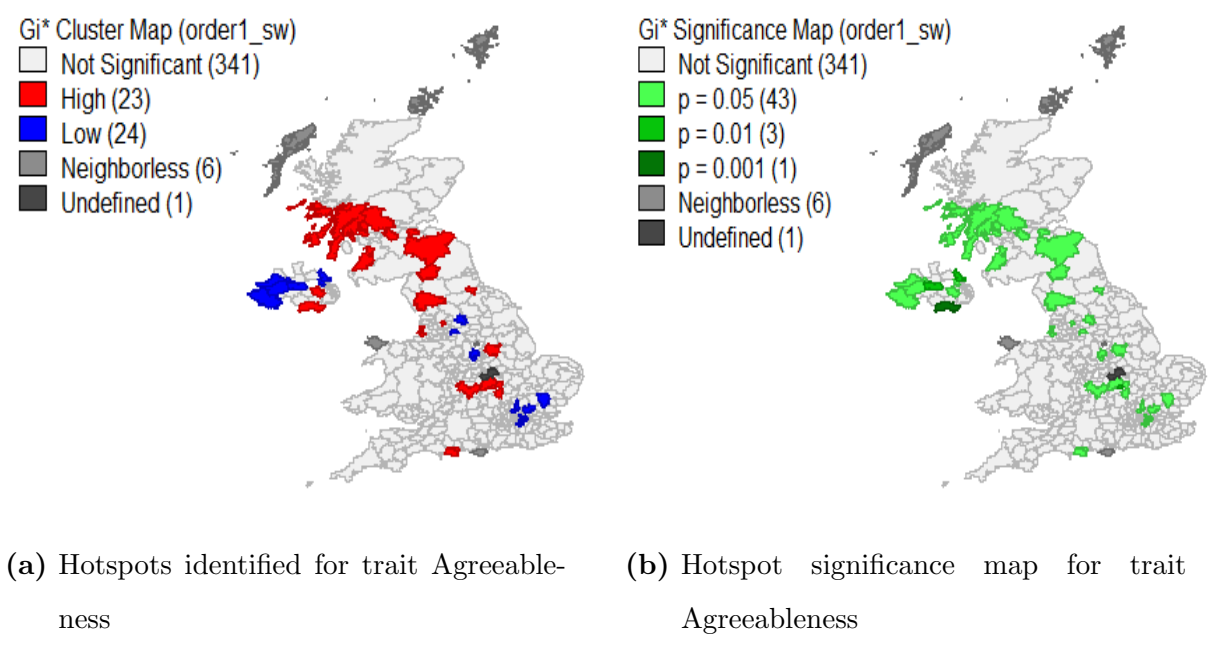


Figure 8.12: Agreeableness

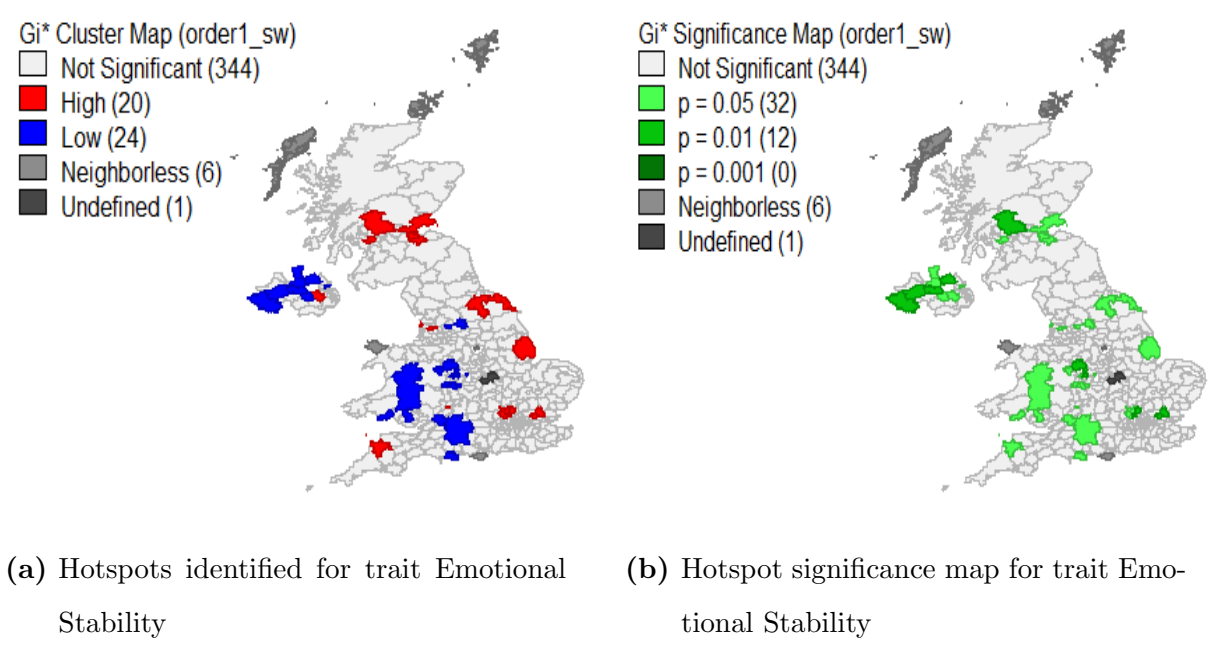


Figure 8.13: ES

8.3.4 Discussion

Some spatial clustering was detected using *Method 2*, aggregating personality traits by district from the stores in which people shop in. The distribution of Boots stores mapped in Figure 8.3 shows that the stores are not evenly distributed. Some districts, which are more urban, have many more Boots stores within them. Whilst rural districts have much fewer. This entails that urban districts have more personality information input compared with rural ones, meaning the validity of the districts' personality may vary from district to district.

The spatial clustering found was, understandably, weaker than that found by (Rentfrow et al., 2015). Rentfrow et al. (2015) used a greater quantity of data on peoples personality, and data where the geo-location attached was the person's residence, not where they shop. Using where people shop at Boots as an indicator of which district a person resides in will inevitably add some noise into the data, as people may cross district boarders to get to their local Boots stores. This is particularly the case in London where the geographic districts are much closer together. In summary, some signal exists for spatial clustering in the UK using the proxy of Boots stores customer's shop at as a proxy of their home district. The next section investigates whether customer personality, aggregated at a store level, is predictive of the store's sales of anti-anxiety medication.

8.4 Data Comparison's

In Section 8.2, regional clustering was studied by aggregating people's personalities, captured from where people shop at Boots. This is likely to be different from where people choose to live, and may not capture the rural/urban divide. Further, people may be buying things from Boots on their lunch break, entailing the measure may be capturing the personality of the places people work. To try and understand better what was being captured when calculating district personality using the customers of Boots store, a final analysis was conducted using 2 different datasets.

The notion that the methods from Section 8.2, capturing where people shop, may produce different regional personality results from where people live, was investigated by comparing the Boots data in London to two other datasets covering London:

- 1) **The BBC personality data**, used by Jokela et al. (2015) and Rentfrow et al. (2015), to see if similar patterns in the distribution of traits were captured.
- 2) **The UK census data**, to check how representative the Boots data is of London boroughs, using demographics.

London was chosen due to there being a high density of individuals in a small area, and lots of close boroughs making it more likely that people will likely travel across these for work.

8.4.1 London Personality Comparison's to the BBC Personality Data

To assess the differences in personality between the London districts personality calculated using the Boots stores, and the London districts' personality captured using peoples residing postcodes, the results from Section 8.2 were compared to the results from Jokela et al. (2015)'s study using 56,019 people from 216 postal districts. This data was from The Big Personality Test run by the BBC and is available online from the UK Data Service upon request. Although the postcode information was not available for download the data had already been aggregated over spatial districts prior to the individual level postcode data being removed. As London is a smaller geographic area than the whole of UK, it was computationally viable to calculate the expected personality of the district. Using this method, if a person shopped more within a district, their personality contributed to the district's personality more than someone who shopped there less often. By visually comparing the spatial patterns it was found that the London borough scores calculated using Boots customers showed a different pattern to that using the data from the BBC personality. These differences are illustrated in figures 8.14 to 8.18.

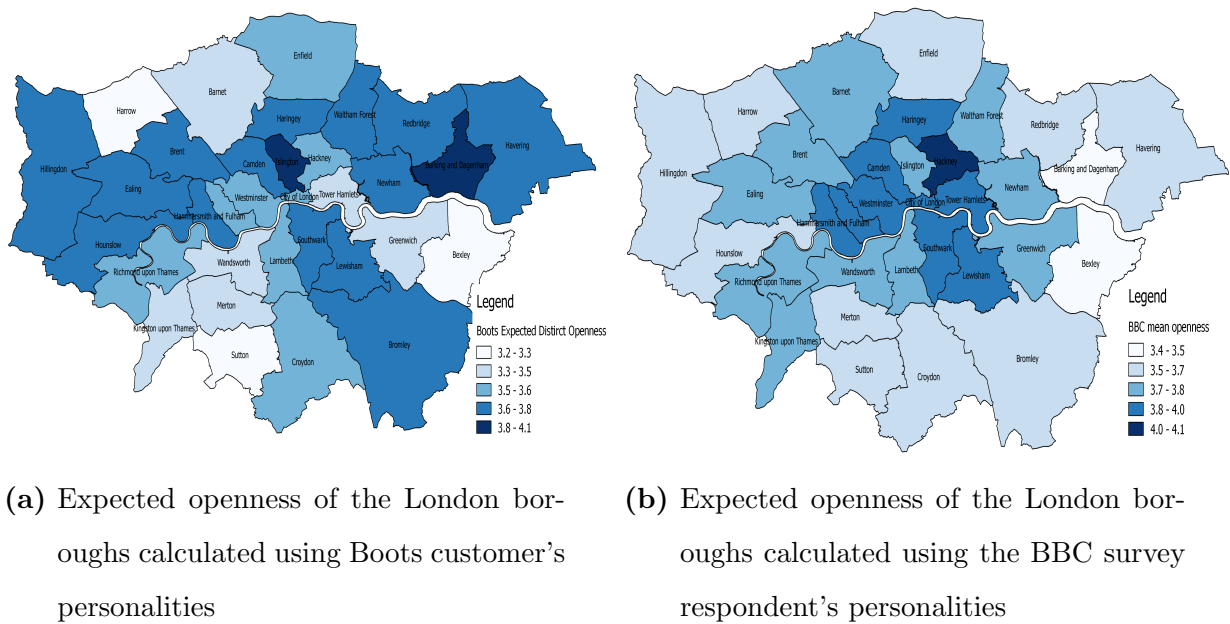
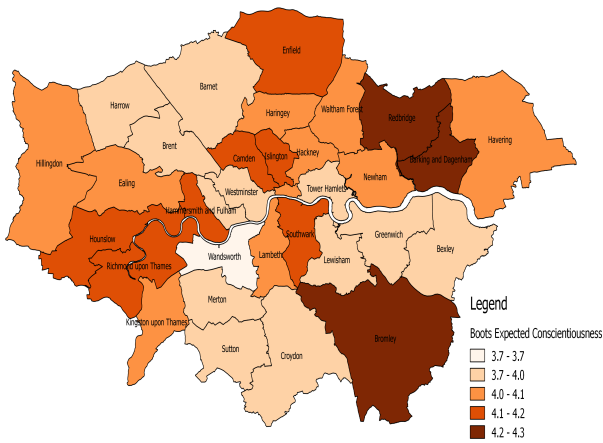
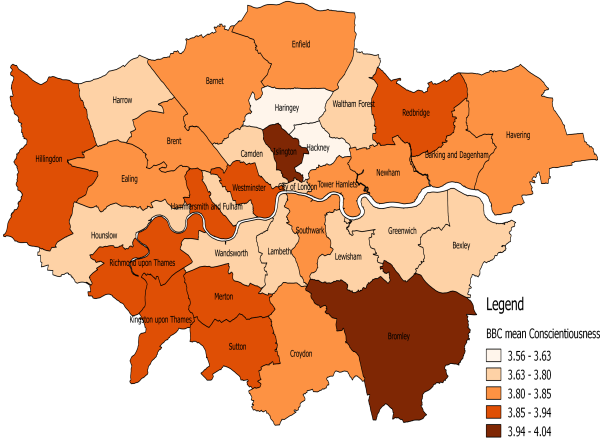


Figure 8.14: A comparison of the expected openness of London boroughs using two different datasets

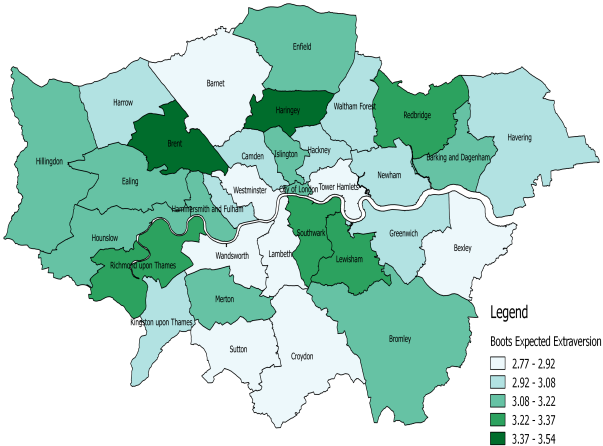


(a) Expected conscientiousness of the London boroughs calculated using Boots customer's personalities

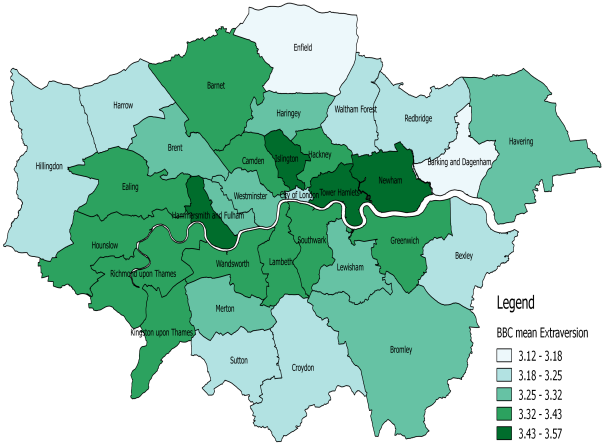


(b) Expected conscientiousness of the London boroughs calculated using the BBC survey respondent's personalities

Figure 8.15: A comparison of the expected conscientiousness of London boroughs using two different datasets

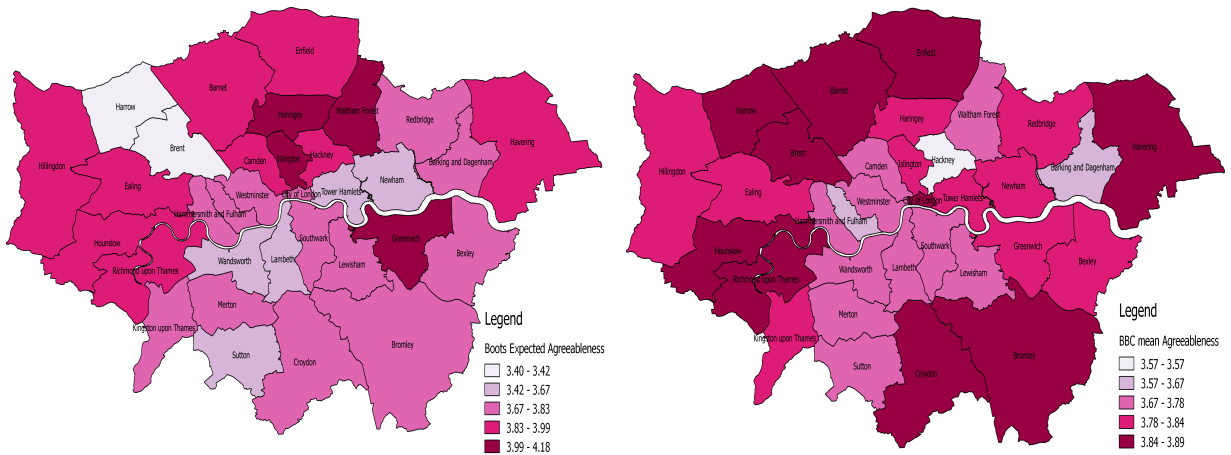


(a) Expected extraversion of the London boroughs calculated using Boots customer's personalities



(b) Expected extraversion of the London boroughs calculated using the BBC survey respondent's personalities

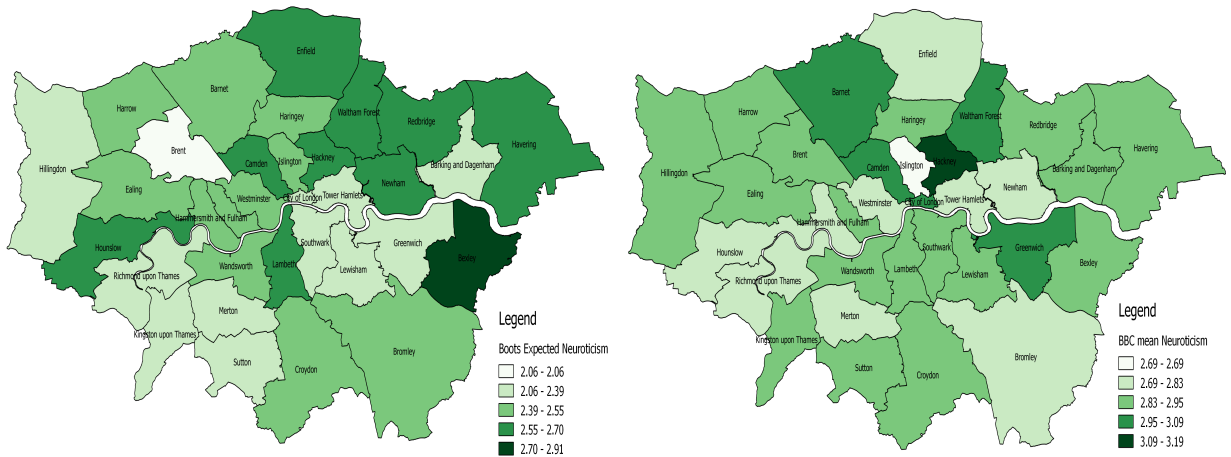
Figure 8.16: A comparison of the expected extraversion of London boroughs using two different datasets



(a) Expected agreeableness of the London boroughs calculated using Boots customer's personalities

(b) Expected agreeableness of the London boroughs calculated using the BBC survey respondent's personalities

Figure 8.17: A comparison of the expected agreeableness of London boroughs using two different datasets



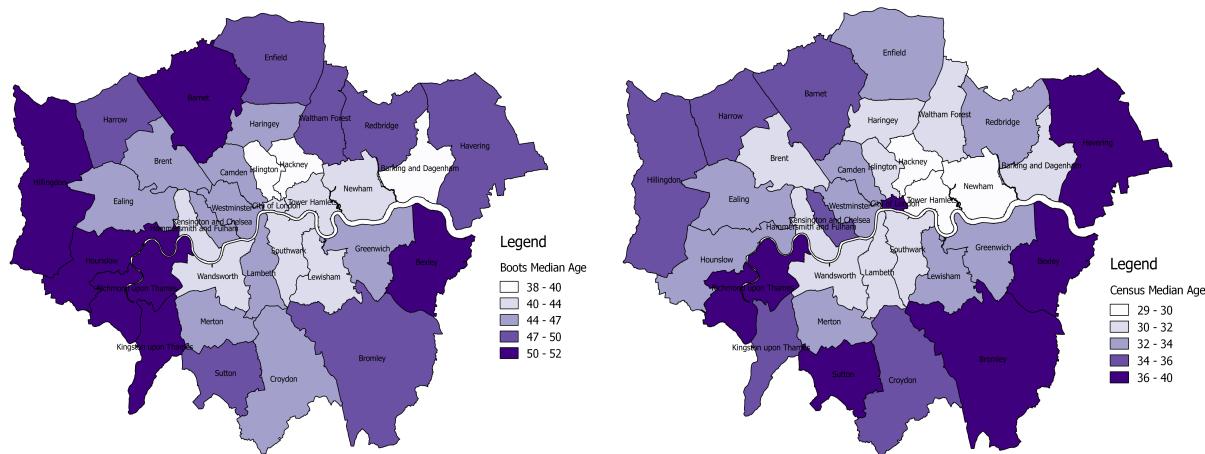
(a) Expected neuroticism of the London boroughs calculated using Boots customer's personalities

(b) Expected neuroticism of the London boroughs calculated using the BBC survey respondent's personalities

Figure 8.18: A comparison of the expected neuroticism of London boroughs using two different datasets

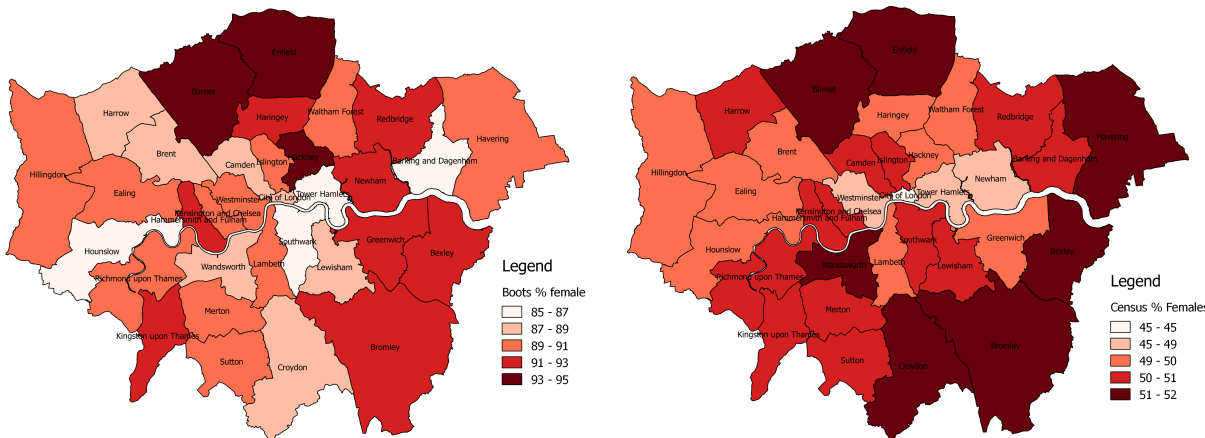
8.4.2 London Demographic Comparisons to the Census Data

After visually assessing that there were differences in the personality of different London boroughs calculated using the Boots stores (compared to using postcodes), a further comparison was undertaken to compare the demographics being captured. The census data for each London borough was compared to the demographics of the Boots customers who's personality was being used to calculate borough's personality. This was to test whether the people in the Boots sample were representative of the actual people residing in the London Borough. Figures 8.19 to 8.21 illustrate these comparisons. They indicate that not only are the Boots customers not representative of the age or gender of the population of the London boroughs, the number of shopping events appears to be negatively related to the population



(a) The medium age of London boroughs from the Boots data. (b) The medium age of London boroughs from the census data.

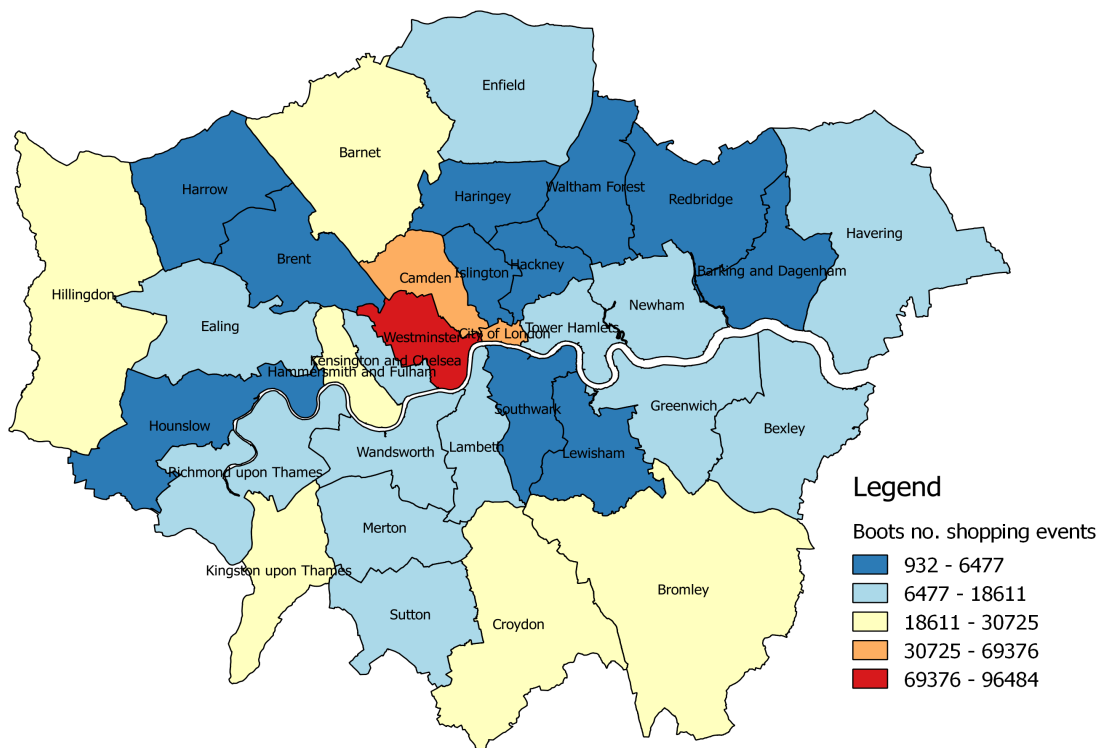
Figure 8.19: Maps of London to compare the age of London boroughs, calculated using different data.



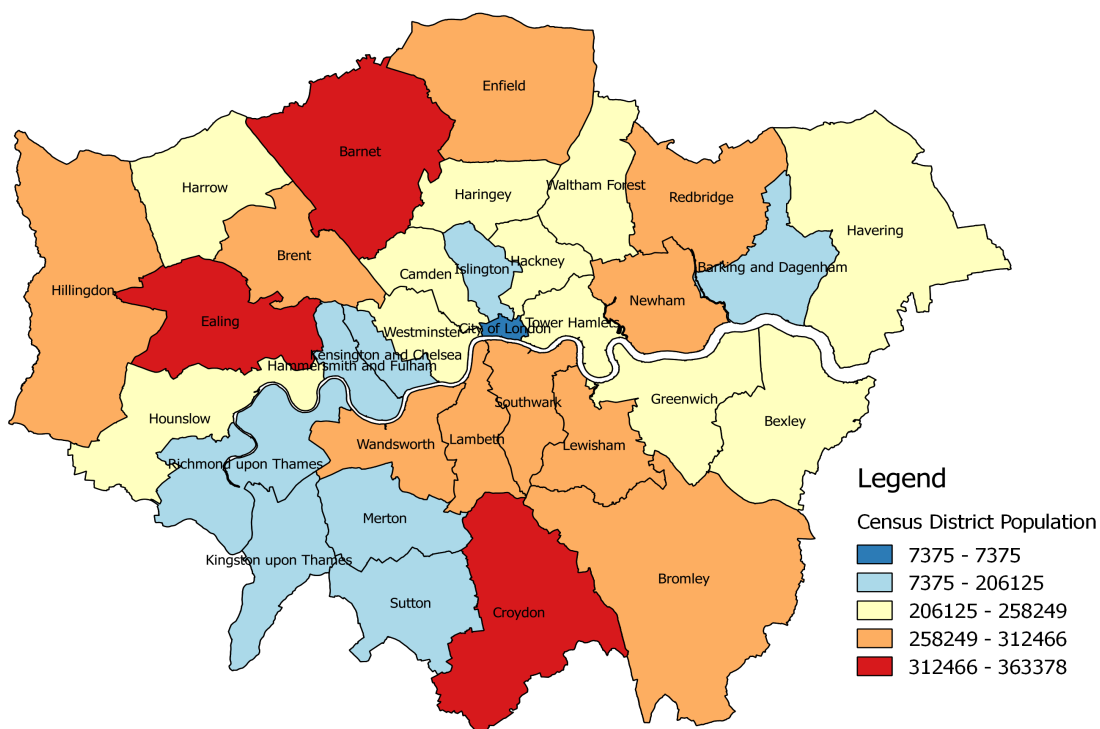
(a) The gender profiles of London boroughs from the Boots data. (b) The gender profiles of London boroughs from the census data.

Figure 8.20: Maps of London to compare the gender distribution across London boroughs, calculated using different data.

of the borough. In other words, the boroughs where people live are not where they shop at Boots. Although these comparison's were only assessed visually, it was apparent that the regional effects calculated using the Boots data could not be interpreted as differences in where people live, and that further investigation would be required before asserting exactly was being captured by the Boots stores, and whether it is useful.



(a) The density distribution of shopping events in the London boroughs calculated from the Boots data.



(b) The density distribution of people in the London boroughs calculated from the census data.

Figure 8.21: Maps of London to compare the density of people across London boroughs, calculated using different data.

8.4.3 Discussion

Geo-tagged big data affords exciting opportunities to study human behaviour geographically. Firstly, it can enable the analysis of much finer grained regional and cultural differences. Secondly, different datasets can be aggregated and joined. This allows for more data to be collated from different sources, possibly collected for different reasons, and potentially

capturing different people. Therefore, the combining of different datasets can allow for the same problem to be studied from different perspectives.

Using data which is geo-tagged has the potential to act as a proxy for home addresses, negating the need to collect and store the sensitive information. However, the data comparisons highlighted that assumptions cannot always be made that where people’s data is being collected indicated the borough or district they live in. This was the case with where people shop at Boots, but it may also be relevant for other data sources such as where people tweet (for most people more waking time is spent at work, than at home) or data on where people run (in the winter people may prefer to run in their lunch hour to maximise on sunlight exposure).

In summary, although the proxy analysed here may be capturing something (such as the area people work/shop), it did not appear to relate to where people live in London. Although acquiring the personality of where people shop isn’t helpful to study the geographical clustering of personalities, it may still be an extremely useful metric for companies like Boots. For example, if Boots know that a particular store is visited more by extroverts than introverts, or people who are more impulsive compared to those who plan ahead, different product ranges or advertising methods could be employed as a result of such information. Thus, in this instance, knowing the personality of the people who purchase from Boots stores, is probably more useful to knowing the personality of the area customers live in.

8.5 Investigating the Geographic Relationship Between Psychological Traits and Anxiety Medication Sales

The results from Section 8.1.2 found that E and N (negatively), and shopping impulsivity (positively) predicted the purchasing of anxiety medication in a regression model. In addition, living in certain regions such as the North East, and Northern Ireland negatively predicted the purchasing of anti-anxiety medication. This may suggest that people in some parts of the UK are less prone to anxiety symptoms. Such an effect could be because of environmental factors, or due to differences in peoples personality, contributing to a lower predisposition to anxiety. Section 8.2 indicated that there is some geographic variation in regional personality traits, and these may reflect personality difference in where people shop. Perhaps these could explain the regional differences in anxiety medication sales?

The relationship between district personality traits and health outcomes have been investigated in the UK. Rentfrow et al. (2015) found that regions high in E were associated with a longer life expectancy, whilst regions high in N were associated with long-term health problems, deaths from stroke, cancer, and heart disease. Yet, the relationship between the prevalent personality traits in an geographic area, and mental health has not previously been studied in the UK. In light of this, this next analysis investigates whether customers personality traits (particularly E, N, and shopping impulsivity) aggregated by *store*, predict the

anxiety sales of the store. Note that here, the store level, rather than district, was chosen due to there being more data points for stores compared to districts; and using a store is more granular (can depict the personality of the people who shop in a particular neighbourhood, rather than spanning across several towns or cities).

For this analysis, the *expected value* for each personality trait per store was calculated (see Equation 8.4 where X_i is an individual's the personality trait). This is similar to *Method 2* as described above, but the store's personality was weighted towards customer's personality who visited more frequently (by the person's number of visits), and can be conceptualised as the best guess of a customer's personality who visits a certain store, given the personality information on (some of) its customer base. This analysis extended the work by (Rentfrow et al., 2015) by calculating the expected store value for other psychological traits (not just the Big Five) such as shopping impulsivity, self control, BAS, as well as happiness and satisfaction to investigate if any of these, aggregated to a store level, relate to the sales of anxiety medication.

$$EV = \sum P(X_i) * X_i \quad (8.4)$$

8.5.1 Results

The geographic distribution of over the counter anxiety medication can be found in Figure 8.22. Correlation analyses were carried out to identify whether any of the expected store level personality variables were related to percentage of anxiety sales in that store. Contrary to predictions, neither trait N ($r = 0.01$, $t(2359) = 0.68$, $p > 0.05$) or trait E ($r = 0.02$, $t(2359) = 0.99$, $p > 0.05$) showed a significant relationship with the anxiety sales by store; yet shopping impulsivity did ($r = -0.08$, $t(2357) = -3.64$, $p < 0.001$) and self control did ($r = -0.06$, $t(2358) = 3.61$, $p < 0.01$). The relationship between shopping impulsivity was the opposite to what was expected, given the regression results earlier in this chapter (that greater shopping impulsivity predicted great purchasing of over the counter anti-anxiety medications). The full correctional results are illustrated in Figure 8.23.

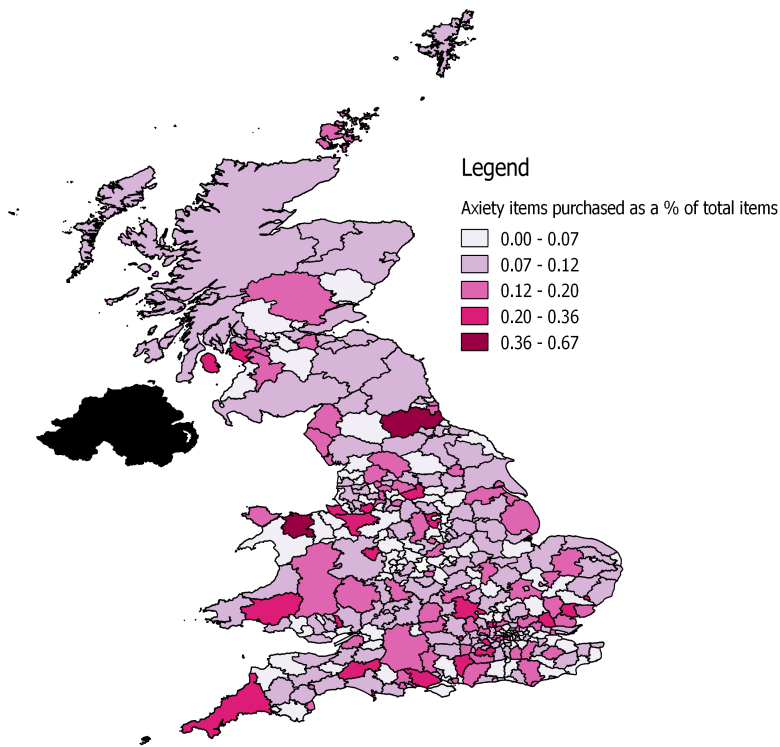


Figure 8.22: Anxiety medication purchased as a percentage of total items purchased, aggregated by district for illustration purposes.

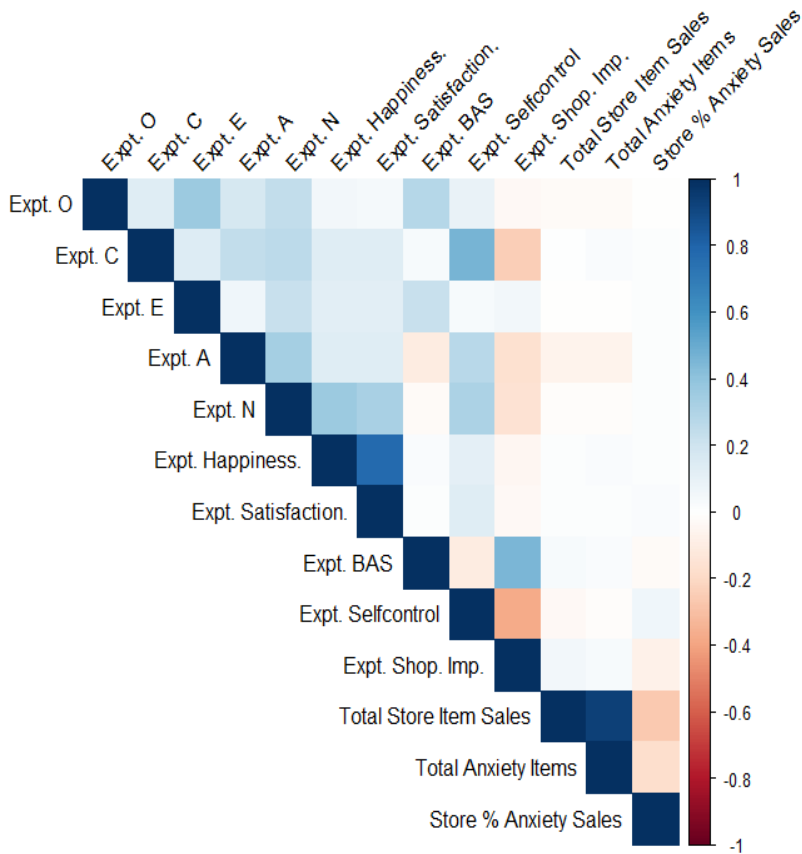


Figure 8.23: The correlations between the expected store’s customers personalities

8.5.2 Discussion

The correlational results show that traits E and N (which were predictive of anxiety sales on an individual level) were not related to sales when aggregated at the store level. Self control positively related to the store anxiety sales, and shopping impulsivity did negatively. Interestingly, the latter effect was the opposite from what was found in the individual level regression analysis in Section 8.1.2. This emphasises that, although many big data psychology research use aggregated data to deduce something about the individual's psychology (see Chapter 2), that effects don't always generalise from an individual to aggregate level, and vice versa (Simpson, 1951).

Self control and shopping impulsivity were found to relate to the percentage of anxiety sales. Although the effects were statistically significant, they were small. This highlights one of the difficulties with working with large datasets (in this case over two thousand stores) - a significant effect doesn't always entail it is a meaningful effect (these issues are discussed in more detail in the following chapter). In this case, looking at a simple bi-variate correlation, with no control variables, it is hard to distinguish signal from noise. For example, store type (or other store information) was not taken into account, and this has potential to influence the sales of anxiety medication. Details such as what products are in stock, or the range of products, will likely introduce noise that can drown out other effects (like the psychological makeup of its customer base). Thus, more store level information would be required to truly delineate the effect of the stores customers' personality on sales.

Chapter 9

Discussion and Conclusions

The previous chapter, highlighted two potential advantages of big data 1) the prospect of employing it to measure psychological constructs (to reduce the reliance on self report measures), and 2) the capture of spatial components that many big datasets have, and how these can be utilised to study behaviour over smaller, more fine grained geographical locations. In addition, a discussion was provided on how geo-tagging can enable the joining of multiple datasets when aggregated over spatial dimensions. This chapter begins by continuing to discuss the potential advantages of big data (as proposed in the introduction), but now in light of the analyses carried out in this thesis. The chapter then focuses on the flip-side, and the limitations of using big data which have presented themselves in this research. Subsequently, the ethical considerations are discussed in relation to big data in general, and directly in regards to big data psychology and the research outlined over the previous chapters. Finally, conclusions are drawn as to where big data psychology should be situated in the prediction versus explanation debate, and the induction versus deduction trade-offs.

9.1 Advantages of using Big Data in Psychological Research

There are many reasons why big data is attractive to researchers in psychology. In Chapter 1 these advantages were generally categorised into four groups: 1) the scale of the datasets. Big data provides access to large samples of real world behaviour quickly and cheaply. This can remove sampling bias in particular an over reliance on WEIRD samples (Henrich et al., 2010), improve the statistical power of studies, permit the use of machine learning algorithms, and allow for more longitudinal and geospatial analyses to be conducted. 2) Capturing behaviour in a naturalistic environment. Big data records behaviour as it is happening in real world situations, removing biases such as demand characteristics which can be observed in an artificial lab environment. 3) Studying the underrepresented. Big data allows us to identify and compare different sub-populations in society that would otherwise be too small to identify and statistically study in smaller datasets. 4) Preventing overfitting. Acquiring enough data to be able to hold some back allows prediction models to be tested on how well

they can generalise out-of-sample, and thus reducing type 1 errors caused by overfitting. In summary, there are many reasons why theoretical researchers in psychology should be excited about the opportunities afforded by big data. The following section discusses the specific advantages and limitations of using big data for psychological investigations, using examples from the work outlined in this thesis.

9.1.1 High Statistical Power

The large sample sizes (and resultant high statistical power) supported by big data allows for smaller effects to be identified and studied. The advantages of this was demonstrated in Chapter 4 which replicated the positive effect of psychological-fit in purchasing data for Subjective Well-Being (SWB) (reported by Matz et al. (2016)), despite the study having small effect sizes. In addition to this, the findings were shown to have generalisability (see Table 1.1) - even when using a different method for calculating psychological-fit (using the data to derive the product and brand personalities to calculate Personality Alignment (PA) in Section 4.3), or using different models (see Section 5.2) the results stood. As the effect predicted was small, such a replication may not have been possible on a smaller dataset due to low statistical power. As pointed out by Boyce et al. (2017) there is however a the risk that findings like Matz et al. (2016)'s, where regressions were fitted to large datasets, run a higher risk of inflating effect sizes (Yarkoni and Westfall, 2017). In addition, the legitimacy of using p -values in large datasets was questioned (see Section 6.5.1). The negative impact of high power in large samples was demonstrated in Section 5.2, where overfitting was shown to have a dramatic inflation effect on the R^2 metric, not to mention p -values. In Section 6.5.1, suggestions were discussed for mitigating against this, such as reporting confidence intervals and effect sizes (instead of just p -values), as well as using Cross Validation (CV) to prevent overfitting. Finally, in Chapter 6, when significance was required to determine important predictors, a method to prevent overfitting using cross validation, and controls for FDR, was proposed.

9.1.2 Big Data Enables the Use of Machine Learning Algorithms

Machine learning algorithms were used throughout this thesis for different reasons. In Section 5.2 the regression model chosen for predicting SWB in Section 4.3 was compared to the performance of other models. The linear regression model performed the same as a random forest, illustrating that there could be an alternative non-linear solution to predicting well-being, different from that being used by the regression model. The non-linearity and interactions were then investigated using decision trees and post-hoc analyses of the non-linear relationships. For example, it was found that personality alignment for trait neuroticism (PA^N) was a non-linear predictor of well-being - it was a strong predictor of those who had low well-being, but not predictive of happier individuals.

In Chapter 6, Section 6.3, an algorithmic approach was used for predicting frequent bag

buyers, selecting a model based on its prediction accuracy. Subsequently, in a refinement of the question in Section 6.4, this method was forgone due to the need for statistical significance of the variable importance measures, and preference for a simple model for interpretation. This trade-off between simplicity and accuracy often depends on the reason why and for whom the research is being conducted. If the prediction model is being used to make decisions in the real world, particularly by those not adept at statistical or algorithmic modelling, the decision maker needs to be able to understand and trust the model they are basing their decision on, thereby favouring simplicity. On the one hand, the studies reported in this thesis have confirmed the continuing value of the logistic regression model, which is easy to interpret (assuming there is no multicollinearity) and produces statistical significance criteria for the predictors. On the other hand, this thesis deliberately compared regressions to decision trees and random forests, which are at the simpler, more interpretable end of the machine learning spectrum. This thesis did not compare logistic regression to other popular machine learning algorithms such as neural networks, where (when optimised) would be expected to perform better than logistic regression (in terms of prediction accuracy). But the case for the use of decision trees specifically is that they can be used to explore non-linearities and subgroups in the data, *aiding* interpretability and future model specification, not hindering it.

An algorithmic modelling approach also allows for multiple models, simple and complex, to be compared, and the robustness of the findings to be tested (see Table 1.1). If the same variables predict strongly across different models, and using different measures of importance, it can be inferred that the effect is robust. If not, i.e. when some models (perhaps non-linear) are more accurate, or produce variant variable importance results, this conveys interesting information about the structure of the data and the relationships between the Independent Variables (IVs) and Dependent Variable (DV). In Chapter 8, although a multiple linear model was pre-chosen to predict anxiety sales, a comparison with other models was also performed. If, for example, a random forest were substantially more successful at predicting over the counter anxiety medication, it would suggest that the linear model was either incomplete or an over simplification - rendering conclusions potentially unreliable.

9.1.3 Sampling Variability

One of the advantages of big data is that it provides the opportunity for greater sampling variability. The demographics of the Boots survey sample, show that it does not predominantly consist of undergraduate students, like many psychology experiments do. Indeed, only a very small percentage of the sample were students (see Figure 3.3). The majority of survey respondents were professionals, with an average age of 51 years. Moreover, there is also a good representation in the survey sample of people from different income groups, different parts of the UK, and different occupation types (please refer to Chapter 3, Section 3.2 for full descriptives of the data).

Despite the data being large enough to statistically study effects within sub-groups (i.e.

bag buying by income group, in Section 6.3) or extract subsets of customers based on their behaviour being extreme/distant from the norm (i.e. identifying the frequent bag buyers in Chapter 6), assumptions about generalisability from the Boots data to the rest of the population need to be considered with care. For example, the Boots data was predominantly females, and a comparison of the survey respondents personality to the scale norms (in Chapter 3) indicated that the respondents have psychological make-up which varies from the norms (this is discussed in more detail in Section 9.2.1). This is important to note as a limitation, as personality has been shown to considerably influence consumer behaviour (see Chapter 3).

Although the Boots data is solely from the UK, the quantity of it enabled statistical comparisons across different countries (see Chapter 6) and regions (see Chapter 8) within the UK to be undergone. Such comparisons uncovered subtle cultural differences (for example, in environmental behaviours and mental health) within an industrialised western country. Furthermore, the utility of large samples, which enable an inductive approach, are useful for the purposes of theory formulation even if the sample is atypical in some way. Such big data analyses can be used to pave the way for follow up, more rigorous investigations (with controls imposed to mitigate sampling bias).

9.1.4 Naturalistic Behaviour Capture

A further potential advantage of big data, outlined in the Introduction, is its ability to capture behaviour directly from the real world (outside of an experimental set-up) whilst remaining inexpensive and resource efficient for researchers. In Chapter 6, it was found that self-reported environmental concern did not relate to the behaviour of buying single use plastic bags. This is an impactful finding as it suggests that the results of studies using self-report behaviour may not be reliable when the findings are generalised to behaviour in the real world. This is particularly important to bear in mind when policy and political decisions are being made in the real world using findings from studies using self-reported data. However, with data like that from Boots, it is important to be aware that additional noise is introduced because customers are not necessarily buying for themselves. In the case of the plastic bags study, this is less important because bag buying is measuring a behaviour about the consumers shopping trip and how they shop. However, for the study predicting anti-anxiety medication in Chapter 8 this adds noise which potentially weakens the prediction signal - because for some records the personality will not be predictive of the behaviour, as the personality information does not represent the person for whom the products are actually bought for.

Even if the individuals are only buying for themselves, the Boots data is of course on its own, still providing an incomplete picture of an individual's consumption habits. For example, a person may buy sleep aids as part of their weekly supermarket shop (as opposed to a Boots store), and thus although we have questionnaire responses of them reporting they

are neurotic and have low subjective well-being (SWB), there is no record of them purchasing anti-anxiety products. Again, this introduces noise into the data. The data could have been reduced to a subset of people who report to be frequent shoppers at Boots, or those whom can be seen from the data are carrying out regular shops. However, this would greatly reduce the size of the dataset. In some cases, greater data quality over quantity can improve prediction models (Faraway and Augustin, 2018). Whether this is the case in the Boots data or not, is something that would have to be systematically tested. Yet such manipulations of the data could be viewed by some in psychology as a form of “procedural *p*-hacking” (Yarkoni and Westfall, 2017) (as it is selecting the population for whom the model works best for). However, if cross validation or out-of-sample testing is indeed used appropriately, this should help to guard against the model over-fitting to a particular subsample of the population (Yarkoni and Westfall, 2017).

Preventing Overfitting

Importantly, and as highlighted in the Introduction, the large quantities of data available for research can enable the invaluable use of hold-out samples and data recycling using CV procedures to prevent overfitting. It also enables the use of machine learning methods which can be optimised to the point in which they *generalise* best, again ensuring models do not overfit. The use of CV as a method to prevent over-fitting is emphasised in the model comparisons of Chapter 5. In this particular example, the linear model was simple enough that it did not overfit (see table 5.1) as the R^2 value was the same for the fitted model and after cross validation. However, the more complex models (the decision tree, random forest, and neural network) demonstrated substantial signs of overfitting, when they were not optimised using out-of-sample data. How cross validation can be used to prevent overfitting is also illustrated by Yarkoni and Westfall (2017) in their simulations. They show that even with simple models, on a variety of different sample sizes, some degree of over-fitting occurs. Big data allows the use of CV and hold-out samples as, unlike experimental data, researchers have data to spare - so generalisability can then be tested without having to sacrifice statistical power.

9.2 Limitations of Big Data in Psychology

The potential for big data to augment psychological research has been reviewed above in relation to the investigations in this thesis. Indeed, the predictive and statistical advantages of big data were demonstrated in Chapters 5 to 7, and additional benefits explored further in Chapter 8, investigating the utility of the spatial properties of big data. Nonetheless, distinct challenges and hurdles for using big data in psychological research were also uncovered throughout the thesis, particularly in relation to the use of statistics on large samples, and data with many features. These challenges were particularly exposed in Section 6.4, when identifying the *statistical* predictors of frequent plastic bag buyers, whilst ensuring “p-hacking

as procedural overfitting” (Yarkoni and Westfall, 2017) was not occurring. Big data is of course not the answer to all psychologists’ generalisability and replicability problems. There is “no free lunch” in machine learning (Wolpert and Macready, 1997), and the following sections discuss the (perhaps less well publicised) limitations of conducting social science research in big data - of which some are still emerging (Faraway and Augustin, 2018).

9.2.1 There Can Still be Sampling Bias in Big Data

The belief that big data removes sampling bias is not strictly true. There is still an element of self-selection required for most people to appear in datasets. For example, a certain type of demographic and personality type will tweet, use Fitbits, or use loyalty cards. Intuitively, it is reasonable to ascertain an over-representation of frugal people, for example, will exist in a loyalty card dataset (due to the savings acquired by the scheme), and these people may behave differently to those who don’t own loyalty cards. This can bring about problems for researchers hoping to study the generalisability of an effect over the population.

Even for datasets like medical records, where data is obtained without individuals signing up to a rewards scheme, there still exists bias. People predominately visit the doctors when they are sick, entailing that information is collected more frequently from people who are regularly sick, or at least regularly believe they are ill (i.e. hypochondriacs). Not only would the data at the population level be biased towards sick people, but on an individual level too. For example, there will likely be less health metrics recorded (i.e. weight, blood pressure, heart rate) from individuals when they are well, compared to when they are sick. In this example, if prediction models were applied to such imbalanced data, without addressing or acknowledging the bias, it could lead GPs to make fatal underestimates of risk in well people, or at the very least make unreliable or inaccurate estimates. Consequently, it is important researchers recognise (and even measure) the representation, biases, and limitations of the data.

Legislation on personal data is predicted to get tighter (Petrov, 2014). As with the introduction of GDPR laws in 2018, this will likely further change the psychological and demographic profiles of those individuals appearing in particular datasets. Consider in medicine Herrett et al. (2015) showed that there was less variability when schemes changed from opt-out to opt-in. In the future, there will likely emerge schemes where people can actively donate their data for research (Skatova et al., 2014; Skatova and Goulding, 2019). Although there is research on the value people place on their data (Skatova et al., 2013), little is known about how people feel about sharing data for public good research (Skatova et al., 2014). Skatova et al. (2014) found two core motivations for people donating data for such purposes: self-benefit, and pro-social concern for others. Yet, despite recent survey investigations (Skatova and Goulding, 2019), less is still understood about what psychological traits might initiate people to actually donate data for different research causes. Such a question may be worth answering in full as the social sciences increasingly employ big data, attracted

by high sampling variability and achieving generalisability.

The descriptives of the Boots data, outlined in Chapter 3, indicate that it may be wrong to simply assume a big data sample’s psychological makeup mirrors that of the general population. As noted the respondents were substantially more Agreeable (A), more Conscientious (C), and less Open to New Experiences (O) than the scale norms described by Gosling et al. (2003). Such differences may be resultant (at least in part) from the Boots respondents being on a volunteer viewpoint panel. However, the differences could also reflect a type of personality which is more willing to donate data (for example, being high in O). Even within the pool of individuals willing to donate data, it would be expected that big data samples’ ‘psychological make-up’ would differ according to how the data was collected, what incentives were offered (if any), whom the data was collected by, and what the data was (said to be) used for (Skatova et al., 2014). Therefore, even in big data there is likely to be some bias in the data, and researchers need to be aware of this.

9.2.2 Machine Learning Algorithms can Discriminate Against Sub-populations

The previous section outlines how, even in big data, the samples may not be representative. The following section outlines the issues which can emerge when machine learning algorithms are used on biased data, without such biases being addressed. Despite a proposed advantage of big data being the ability to study the under-represented in society, machine learning algorithms can be used to actually worsen the marginalisation of minority groups. Several cases have come to light where algorithms used in society have been exposed as being unfair or biased towards a particular part of the population. For example, COMPAS is an algorithm being utilised in the US to predict the likelihood of an individual re-offending. It is being used to guide the severity of a sentence given. In 2016, it was reported that the algorithm over predicted the likelihood of re-offending for black defendants and under predicted for white (Cossins, 2018). A study by Dressel and Farid (2018) found that the algorithm in use, performed no better than random untrained individuals making decisions online. This has serious implications, as the mechanisms which may be causing the reported racial bias cannot be independently scrutinised, as the algorithm is proprietary.

Unfortunately, this is not the only case of this nature, more and more cases are emerging of companies and government departments using these inherently biased algorithms (Cossins, 2018). An algorithm called PredPol, currently being used in the US, is designed to predict the times and locations that crimes will occur, aiming to help police distribute resources more efficiently. However, PredPol is sending officers into neighbourhoods where people of ethnic minorities live, irrespective of crime rates (Cossins, 2018). Ensign et al. (2017) showed that the algorithm was learning from police records, not actual crime rates. Thus, rather than statistics, human prejudice was being learnt, creating a “feedback loop” unfortunately compounding the existing racial bias that exists for many police officers (Cossins, 2018).

Learning from such cases is crucial for the development of fair and unbiased algorithms. In this case, it was the type and quality of input data that was causing the bias.

It is not only ethnic minorities which are being discriminated against by algorithms. A 2015 study found that 11% of the images displayed on Google, when the term “CEO” was searched for were women, despite women making up 27% of American CEO’s. The algorithm has a bias towards males for positions of power, and over-represents them in search results. Gender biases are also filtering into advertising. Datta et al. (2015) showed that Google’s online advertising showed more jobs with a higher income to men, than to women. Even if these biases are present in human beings, perhaps we need to start thinking about how we can collect and use data/machine learning to help correct them, rather than exacerbate the problem. Cognitive biases such as the mere exposure effect (Zajonc, 2001) and familiarity bias (Foad, 2010) can powerfully influence how humans perceive and feel about things, given their amount of exposure to them. Consequently, we should seriously consider what information we are exposing individuals to, and the effect it could have on them and society.

Many would argue that academic researchers have a duty to protect individuals from harm, and use research for public good. This includes preventing minority groups from being discriminated against. The majority of algorithms shown to be biased have been created by private companies, which make a profit from selling them to the state, or to other companies. Being developed by private companies makes it harder for researchers to interrogate the algorithms’ fairness, and the mechanisms behind it. Many academics are fighting back against this. Open software tools, such as Equitas (Saleiro et al., 2018), are being developed by universities, for use by machine learning developers, analysts, and policymakers to evaluate the fairness of models and algorithms. In many cases, it is not the models and algorithms themselves that are biased, it is the data being use. Researchers using big data, therefore, should be open about potential biases in their samples, and show caution with the conclusions they draw on generalisability.

In the Boots data, there is a clear bias towards women, and thus the conclusions are limited in how well they can generalise to men. To somewhat mitigate this, gender was always included in models as a control variable. The most dangerous type of bias is when it is unknown or unmeasured (Ghani, 2019). When interrogated, algorithms have been found to be biased against particular ethnic groups, races, and women because there are *labels* for these in the data. Labels (i.e. for race) mean that the biases can be tested for these identified sub-groups or sub-populations. But what happens when such labels don’t exist? How can we test for bias against unidentified sub-populations, for example personality types? This research was quite unique in that it measured the personality traits of people in the data, and therefore had the opportunity to identify bias towards a particular personality type. Yet this is something rarely seen, despite having real world consequences. For example, an intervention may only attract or work for people of a particular psychology. This highlights the potential utility of surveys/data collection polls including short personality measures, along

side demographic questions, to better understand possible psychological biases in datasets.

Understanding psychological bias in data is particularly pertinent in health research, where personality has been shown to have a strong effect on many health outcomes and health behaviours. For example, Neuroticism (N), described as a trait of “profound public health significance”, is a robust correlate and predictor of many mental and physical disorders (Lahey, 2009). Further, C has been shown to heavily influence the adherence to medication (Molloy et al., 2013). If we move towards personalised health care, or data driven early warning mechanisms, for example using the purchasing of over the counter anti-bloating medication (for a certain demographic) to predict risk of Ovarian cancer (Yawn et al., 2004), or the purchasing of sleep aids as a precursor of more prominent state anxiety (as explored in Chapter 8), then we need to expose the ways in which such health data may be biased, to prevent potentially harmful consequences. To illustrate, recommending group therapy to an anxious introvert may exacerbate their symptoms, not treat them. As novel technologies get trialled, researchers and practitioners should bare in mind that they may be being trialled predominately by people high in O and Extraversion (E), as people of these personalities are more likely to seek new and novel experiences. Subsequently, the engagement in, or effects of, new treatments may not easily transfer to introvert individuals who dislike change.

To summarise, researchers using big data and machine learning, particularly those in health, may want to consider the possibility of measuring personality to mitigate against unmeasured psychological bias in the data. Once bias is known and measured, researchers can statistically control for, or eliminate it, if they see necessary (i.e. by extracting ‘unbiased’ sub-samples). At the very least, the potential for psychological bias to occur should be acknowledged - and caution, rather than blind confidence, undertaken in asserting the generalisability of large datasets, to help prevent psychological discrimination of algorithms.

9.2.3 Traditional Statistics for Quantifying “Uncertainty” are Not Useful on Big Data

A reoccurring tension throughout this thesis, and the submission of research to academic (psychology) journals, was between the ‘need’ for statistics (from a psychology audience), and the statistical consensus that in large samples inferential and Bayesian statistics are unreliable at best, and completely useless at worst. Thus, for those looking to quantify uncertainty (i.e. of the effect generalising outside of the sampled population), that statistics “don’t work well” on big data is considered a serious limitation (Faraway and Augustin, 2018). On big data, all usual statistical tests conducted in psychology become significant when the sample is large enough (Lallich et al., 2006). As discussed in Chapter 6 (Section 6.5.1), for any finite parametric model on big data, confidence intervals are exceptionally narrow and p -values are exceedingly low. For Bayesian analyses, the prior gets swamped, and the estimates derived from Bayesian inference are far too certain than common sense would suggest (Faraway and Augustin, 2018).

Some researchers in statistics have tried to use work-arounds, creating “test values” to standardise the effects (Lebart et al., 1995). For example, Morineau et al. (2006), introduced a novel criterion called the TV100, a test value calculated to produce estimates, as if the results had come from a sample size of 100. Much of this work, however, is orthogonal to the field of machine learning where they tend to avoid such issues by not attempting to measure statistical uncertainty at all, instead focusing on prediction accuracy.

Prediction accuracy can be viewed as a type of uncertainty measurement for a model. If the accuracy is low the algorithm can be interpreted as being uncertain on how to predict the outcome. However, what about when we want to quantify the uncertainty of a particular IV’s relationship with the DV (like CIs and p -values)? Such uncertainty could be gauged using permutations to measure variable importance; but this is hard to standardise and compare across multiple models and datasets (which contain differing amounts of variance).

Furthermore, uncertainty doesn’t just arise from the model parameters, as captured in CIs and p -values, different types of uncertainty can emerge from the data itself. There is uncertainty about potential bias in the data (particularly when the researcher is removed from the data collection process, as is often the case with big data). There is also uncertainty around which model is best to use, particularly if models are unstable or the metric used to evaluate them changes how they comparatively perform. And, hence, there is also uncertainty around which metric is best to evaluate the model(s) on. If all this uncertainty could be incorporated into algorithmic modelling, a more realistic picture of the performance and the conclusions could be generated - however difficult this may be to achieve (Faraway and Augustin, 2018; Strobl and Zeileis, 2008).

Models built on small data are simpler (out of necessity) but do often attempt to measure uncertainty using statistical significance. Small data models cannot portray uncertainty perfectly - the mis-specification of models can produce biased coefficients and error terms. Nonetheless, uncertainty is reported and talked about in small data analyses. Such conversations force both the user and the readers to realise that there is some possibility that the conclusions are wrong. In big data, when machine learning methods do not attempt to estimate uncertainty, the notion it exists at all could be entirely overlooked. This is particularly harmful when the large number of variables measured in big data promotes the testing of a greater number of hypotheses - increasing the risk of false discovery. Although a strict CV regime can help prevent the model (in its entirety) being wrongly found to be useful, when analysing variable importance, CV is less useful for mitigating the false discover rate (FDR) for individual features (which commonly requires the model is fitted to the data). Thus, some statisticians suggest that at the very least FDRs should be controlled for - and now a number of researchers are building such controls into their algorithms (Peña, 2008; Lallich et al., 2006; Armen and Tsamardinos, 2011; Nechay et al., 2009; Li and Wang, 2009).

As machine learning algorithms continue to be found to exhibit bias (as described in Section 9.2.2) there is a growing need for transparency and judgements of uncertainty. There is growing momentum in recent research towards uncertainty quantification in machine learn-

ing, reflecting the growing need for algorithms and their developers to be held accountable if algorithmic discrimination is detected. A recently developed method called Bayesian Deep Learning (Kendall and Gal, 2017) is being used to quantify uncertainty in black box models, particularly in computer vision. If a model is uncertain about an image, it can be flagged and additional ground truth labels can be manually added. This also supports more efficient data collection (as data need only be collected in uncertain areas). Uncertainty development will also improve algorithm safety, for example in the context of driver-less cars. Research has also looked at methods for the statistical representations of variable importance rankings, and use of p -values for feature selection procedures (Paul et al., 2013). While this field is still in its infancy, the movement towards quantifying uncertainty in big data is an important one, and, is preferable to falsely assuming certainty (Faraway and Augustin, 2018).

9.2.4 Visualising and Checking Data is Harder with Big Data

A final difficulty which arises when dealing with big data, compared with small data, is that it is hard to visualise. This difficulty is illustrated in Figure 9.1. Visualisation can be an important part of preventing model mis-specification. If the relationships between variables can be visualised, there is less chance of fitting the wrong model, particularly with regards to linearity. When relationships can't be visualised, using an algorithmic modelling approach and selecting a model based on performance, mitigates to some extent the issues that can occur when the structure of the data is assumed. Nonetheless, in the case of complex models, there is a large number of different ways to construct a model, what with parameters, meta-parameters, and variable transformations. Using a skilled eye to detect outliers, interactions, and non-linearities can prove a more time and computationally efficient method for narrowing down these options. To accommodate this need, more data visualisation tools and software are now being developed for use on big data. These include Tableau, Google chart, and D3.js among others.

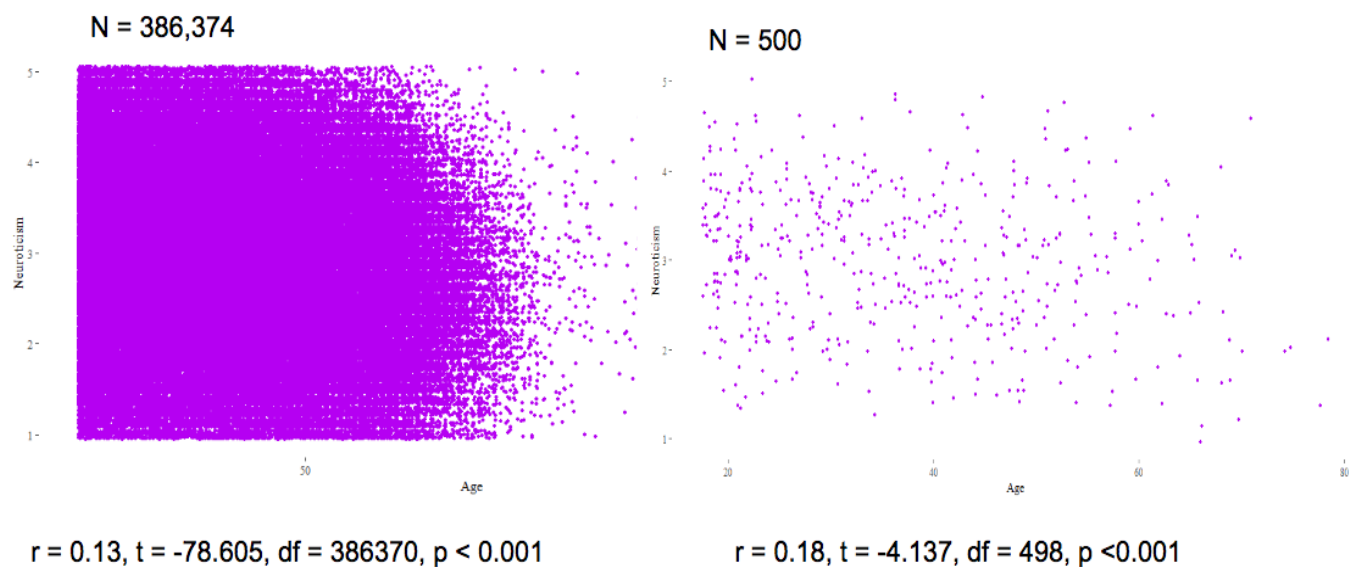


Figure 9.1: Graph to illustrate the difficulties of visualising effects in big data. This data is from the BBC Big Personality Test. The left graph is the whole sample, and the right graph is a randomly selected sample of 500 individuals. On the x axis is age (in years), and on the y axis is Neuroticism. The statistics at the bottom are from a simple linear regression model.

In addition to visualising data for model specification and selection, *viewing* the data can be an important part of data cleaning and sanity checks. Data cleaning is a vital part of big data research, as data can easily be mis-recorded by both humans and machines. For example, there can be bugs in a system which affect data collection/generation in a particular way, and these need to be accounted for. Being able to view the data enables obvious outliers or erroneous data to be detected. This helps to ensure greater data quality and less bias, as well as checking the face validity of the data generation methods, i.e. “Is it measuring what I think it should be measuring”. Thus, it should be noted that one of the downsides of working with big data, is it can be tricky (but not impossible) to visually check and analyse it.

9.2.5 Summary

Despite all the potential advantages of big data, as described in Chapter 1, “bigger” is not always “better”, as such data may still not be representative. Faraway and Augustin (2018) use the crude example of the prediction of the 1936 US election to emphasise this point. Literary Digest magazine ran a poll of 2.4 million people, which predicted (57%) that Alfred Landon would beat Franklin Roosevelt. When the ballot papers were counted they showed a landslide for Roosevelt with 62% of the votes. Meanwhile, a much smaller (only a few thousand people) but more representative poll from George Gallop, predicted a Roosevelt victory with 56%. This illustrates that collecting more data isn’t superior if the additional data is generated from the same mechanism or sub-population (Freedman et al., 1998). The problem in today’s big data samples, is that the bias is not always obvious or measurable, but still has the potential for devastating consequences. Thus, ‘small’(er) data samples which

have been statistically sampled to the population means are preferable to big data that is biased.

Because the size of data doesn't *inherently* make it better, other metrics or frameworks may be required, and need to be developed, in order to capture the elements that make big data valuable for researchers, for example: representativeness, a lack of bias, the number of control variables present, the number of demographic/psycho-graphic labels, as well as the trustworthiness of the data source and data generation process. As more big datasets are being opened up for academic research, such frameworks and ratings of data quality would not only be incredibly useful for researchers, but would also help to ensure the transparency of the potential limitations of the data/findings' generalisability. Overall, big data indeed enables a greater certainty that the results will generalise to the population from which the data were drawn from. Yet, it is important to realise that this population may well be biased, and thus results can not necessarily generalise to a wider population.

In summary, big data offers tremendous potential to supplement psychological research conducted on small, experimental data. Big data can be used to obtain objective and unobtrusive behavioural measures of large amounts of people from different walks of life. However, big data will often come at a cost of some sort, whether it be acquisitional, computational, or privacy costs (Faraway and Augustin, 2018). Further, big data is often noisy and can be poor quality. Thus, small data and big data can both have their place in psychology. But when a causal inference is desired, quality beats quantity in data (Faraway and Augustin, 2018).

9.3 A Discussion on Effect Sizes in Big Data Psychology

Many of the effect sizes in this research were smaller than those found previously in the literature using small data. This, is symptomatic of the noise in big data (Matz et al., 2017) and the temporal disconnect between the time in which the surveys were answered and when the shopping occurred. The latter point should be seen as an advantage, as it prevents well established biases (such as consistency motive, co-variation bias, or common method variance) which can occur when two variables are measured on the same survey, or at the same point in time. Thus, perhaps the effect sizes in big data, where the aforementioned biases don't occur, are in fact more realistic than those reported using surveys or questionnaires.

Large effect sizes are preferable in the literature, as they are more likely to be replicated (Open Science Collaboration, 2015). Yet, the field is still waiting to see how replicable big data psychology studies are. Those who are researching in big data are still interested in what novel findings the data can bring about, and therefore not enough replication attempts have been made to analyse the general (context non-specific) reproducibility rate. Once a greater number of datasets are opened up for research purposes, researchers will be able to

ascertain whether the small effects that are being discovered now in big data are reliable and replicable across different data types and sources. Until then, researchers can only ensure that findings are reliable within the dataset by ensuring the generalisability of findings using cross validation and out-of-sample testing.

Even when generalisability of models within a dataset is strong, small effects could just be an artefact of the noise in the data, and not generalisable to other datasets. In the cases when they do generalise to other data, when are effect sizes considered so small that they are rendered not useful or not meaningful? Cortina and Landis (2009) suggest that even small effects can suggest strong support for a given phenomenon if they occur in the context of inauspicious designs. Given this, it could be argued that given the noise in the Boots dataset - where many customers are expected to be buying items for other people, and the data represents incomplete consumption patterns (as people don't always use their loyalty card, or they buy things in other shops) - that it is amazing there is any signal linking psychological variables to consumer behaviour.

Other instances when, even very small effects, can be meaningful is if the finding challenges theoretical assumptions; or if they have large cumulative consequences. An example of the latter is the Abelson's Paradox (Abelson, 1985) in baseball. The historical batting performance for a player, when they step up to bat is so small (less than 0.5%) that it is rendered useless for predicting their performance. Yet, the cumulative effect over the hundreds of bats played in a season yields substantial differences (of hundreds of thousands of dollars) in batter's salaries. Thus, the impact of many incremental small effects accumulating over time, or over many individuals, should not be so easily dismissed.

When designing the questionnaire for this research, it was predicted that The Big Five personality traits would be strongly predictive of a variety of different consumer behaviours. Despite being strong predictors of well-being (see Chapter 4), the traits did not have as large effect sizes as expected. A potential reason for this, could be that the questionnaire only measured the coarse trait level, and only in 10 questions (2 questions per trait). This was to reduce the length of the questionnaire and consequently the expected drop-out rate. However, Some traits, particularly openness, have different facet levels (intellect and openness) (DeYoung et al., 2009), where, although the facets are related, different facets can predict different behaviours. If one facet strongly predicted a behaviour, and the other did in the opposite direction, then the effect is effectively halved. Thus, with a longer personality test, we would expect larger effect sizes, as well as a more detailed facet-level understanding.

Effect sizes must also be evaluated within the context of what they are measuring. Complex psychological phenomena like SWB, in the real world, are unlikely to be explainable by a handful of strong predictors. This is one of the advantages of big data, that it can be used to unravel the weaker, more subtle predictors, that small data cannot. For example, the effect size of PA on well-being in Chapter 4 is small, but theoretically it should be a small effect, particularly compared to variables which are known to have a strong effect such as emotional stability or income. Indeed, it would be concerning if consumption had a

comparatively larger effect on SWB than what was found.

Overall, small effects found in big data should not be contrasted to those found in experimental psychology studies, as they are simply not comparable (noted above as a limitation of using big data). Effect size comparisons between big datasets may also not be valid, as the noise and quality of the data will vary dramatically between such datasets. Thus, careful consideration of the effects in big data should be made and always evaluated in relation to their context and the data.

9.4 A Discussion on the Limitations of using Survey Data

Having discussed, in depth, the limitations of big data (in reference to the Boots' transaction data) it seems pertinent to also consider the limitations of the survey data in this research, and more widely in psychology. In recent years, in the context of a heightened focus on the replicability crisis (Nosek et al., 2012; Open Science Collaboration, 2015; Shrout and Rodgers, 2018; Zwaan et al., 2018), a lot of attention has been attributed to the data analysis methods used (promoting more stringent significance testing, or bayesian approaches). Similarly, this thesis has focused a great deal on how big data affords new methods (such as CV, out-of-sample testing, and machine learning algorithms) which have the potential to help with the replicability crisis in psychology. Nonetheless, such methods are only useful if the data is of good enough quality; and with all the focus on methods and sample sizes, few have studied the contribution of dubious data generation to the replicability crisis.

Compared with other scientific disciplines, psychologists encounter unique challenges. As a subject matter, psychological and social processes are highly complex, intangible, influenced by many things, irreversible (due to people's memories and experiences), and only characterized by weak laws. Thus, psychology is of a fundamentally different quality to harder sciences like Physics or Chemistry. Yet, perhaps in an attempt to make psychological research more empirical, there has been a trend away from measuring behaviour in psychology (Baumeister et al., 2007). Instead, researchers have preferred using recorded responses (i.e. clicks or agreement ratings), which are intrinsically more easily quantifiable.

Despite their popularity, increasingly research is finding fault with data collected from ratings. For example, using popular scales to measure the Big Five personality traits, correlations between questionnaire items were found to vary unsystematically across different countries (Ludeke and Larsen, 2017). Additionally, unexpected differences in personality between students and the general population have been found (Hanel and Vione, 2016). Findings like these have lead many to seriously question what these ratings are actually capturing.

For different people, in different contexts or cultures, "agree" on a likert scale can mean different things. Uher (2018) showed that *nine* possible reasons could be given for select-

ing “agree” on a likert scale. Even if agree always meant the same for everyone, in what way does the word agree relate to a number or quantity? When the scales are quantified, psychologists are imposing that agree is equi-distant from “strongly agree” and “neither agree nor disagree”, and that this is the same for everyone. A likert scale cannot capture whether someone who selects “agree” is usually a reluctant agree-er or not, how easy it was for them to answer the question, or how confident or certain they are with their response. Consequently, not only are the scales limited by what they can capture, both their design and quantification can be tainted by the egocentric and ethnocentric biases of the researcher (Uher et al., 2013).

Large numbers of respondents, like those gained in this research, can mitigate against some of the reliability issues which can arise when using rating scales. Statistically, the inter-individual variance in how participants use the scale, is more likely to be “washed out” with large numbers. Nonetheless, despite having nearly 15,000 questionnaire responses, the limitations of trying to capture the intricacies of personality, using 10 answers to questions on a 7 point agreement scale, still remain. In the future, big data may be able to help reduce the over-reliance on survey data for measuring both behaviour and personality. As discussed in Chapter 8, Gladstone et al. (2019) have shown that personality can be predicted using people’s consumer behaviour, captured in digital traces. As such predictions improve, using algorithmic modelling and machine learning, behavioural traces in personal data associated with a particular personality, can be exploited as a proxy for the individual’s psychological profile. Thus people’s personalities could be inferred from their data - reducing the reliance on introspective rating scales for measurement.

9.5 Ethics and Big Data

This thesis has focused on the potential advantages that big data can bring to the academic discipline of psychology, and indeed the wider research community. Yet, big data psychology exists within a wider context of increasing use and reliance on big data and AI in society. As well as the fruitful opportunities to enhance people’s lives, the AI movement is also accompanied by growing concerns on the potential negative impacts, from which academic big data research is not immune. Thus, before this chapter makes concluding remarks on the discussions engaged in throughout this thesis (such as prediction versus explanation, and induction versus deduction) this section briefly outlines the ethical considerations surrounding big data psychology. The next section therefore introduces, first, the big data and ethics landscape; subsequently, it explores the ethical implications of big data research more broadly, before considering ethics in relation to big data psychology and indeed this thesis.

9.5.1 Ethics in a ‘Big Data Society’

The explosion of ‘datafication’ in modern life has the potential to bring many benefits for society, including more personalised services, health monitoring, fraud detection, and more

efficient use of resources. However the growing quantities of data also pose an inherent threat to personal and national security. With this movement also brings ethical questions, for example, who the data belongs to, and what responsibilities different players have – moral and legal. The scale and fast stream of the data poses problems for governance. It's estimated that every day humans generate 2.5 quintillion bytes of data (IBM, 2011). The increase in data collected is accelerated by the growth of the Internet of Things, estimated at 26.66 billion connected objects in 2019, and to grow to 75 billion by 2025 (a five-fold increase in ten years) (Forbes, 2016). Data was described (and subsequently frequently repeated) in an interview as “the new oil of the internet” by Alex Pentland, Toshiba professor of Media Arts and Sciences at MIT (Petrov, 2014). It's highly valuable, and those who have it in abundance can profit. However, its potential power and future consequences are not fully known. Pentland (2014): “It's a new asset class, a new value, a new money. And we don't have the regulations to treat it like the value class it is” (Petrov, 2014).

Data regulations are indeed getting stricter. The 2018 EU General Data Protection Regulation (GDPR) was in response to a fundamental shift in the way people view their data rights. It is likely that future regulations will involve further strengthening of consent conditions, and even move towards more personalised data privacy plans to account for individual differences in people's privacy preferences. After all, algorithms are here to stay, and can be expected to be an increasingly greater part of our future lives. Many considerations and discussions will continue to emerge as algorithms aid, and even replace, human decision making in some domains. For example, how to program fairness, who/what to attribute accountability to when algorithms make mistakes, and whether to train them to emulate human decisions or try to ‘improve’ them? In potentially life changing decisions, even if an algorithm produces less error, will human error (i.e. from a jury) be easier to accept than algorithmic error? Furthermore, algorithms modelled on human behaviour will also model the biases humans have. How will we prevent ourselves from entering into a closed loop of erroneous biased decisions? These are just a snapshot of the potential ethical issues which big data may yet bring upon us - “we have seen the benefit of big data, but we are only now also realizing the extent of the associated damage” (Faraway and Augustin, 2018).

9.5.2 Ethics and Big Data Research

Researchers require access to data, yet many data sources are owned by large corporations such as Google, Amazon, or Facebook. With mass datasets (especially in combination with one another) comes knowledge and thus great power. The power is truly in the data and not in the algorithms. If an academic were to design an algorithm which was technically superior to those used at Google, for example, its practical impact would be comparatively limited, without the opportunity to train it on the large amounts of rich data owned by large companies like Google. Unfortunately, a company's primary interest is not always public good, and consequently companies making their data open for public good research

is not common. Many companies restrict their data for economic, political, or privacy reasons. However, more companies are now seeing the value in volunteering data for, and even funding, academic research. Facilitating research can help promote favourable brand associations, and increase brand exposure in the media - “allowing companies to do well by doing good” (Glaeser et al., 2018). Nonetheless, researchers fortunate enough to work with big data, may still find that the company requires some control over what findings are reported, or want to steer the conclusions to make the company appear favourable. This would be at the detriment to scientific process, as other academics can not collect the data to replicate the findings. Thus, as well as ethical considerations there is threat to scientific rigour brought about by big data research.

Even for researchers with the best intentions, conducting research ethically in big data can be a challenge. According to Kosinski and Behrend (2017), there is a lack of clear guidelines on the protocols and design of big data research. In addition, researchers lack of expertise and help with storing data securely, and no clear interpretable guidelines on consent for information exists in relation to donated data (Kosinski et al., 2016). Many researchers use open source data, scraped from websites or social media. Such a methodology has been likened to archival research (akin to that conducted in history or literature) which rarely involves participant protection (Bruckman, 2002; Herring, 1996). But, just because something is practically possible (i.e. it gets approved by an ethics committee), does not necessarily mean it should be done. The line drawn between public and private spheres isn’t whether something is accessible or not (Schultze and Mason, 2012; Kosinski and Behrend, 2017). Guidelines should be set out for researchers in order to retain public trust in big data research, and encourage data donation.

However, it is not just researchers for whom this concerns: who is regulating companies and corporations? One of the arguments researchers have for operating in an ethically delicate space, is the need to better understand what companies are doing with our data, and what is achievable, to protect the public by better informing the creation of regulation protocols and procedures. An example of this, is the need to better understand psychological targeting and persuasion, so we are better able to deal with companies like Cambridge Analytica in the future.

9.5.3 Big Data Psychology Ethics

This research not only utilised customer loyalty card data, but also demographic and psychological data. Personality data is sensitive data, particularly, self report measures on an individual item level (before the scores are aggregated). Depending on the measure, questionnaire items can ask revealing questions about people’s feelings of anger, jealousy, laziness, mood stability, social relationships, as well as participation in taboo experiences (such as drug taking and risky sexual behaviour). Not only can the questions themselves be intimate, but the insights revealed can be too. The aggregated trait scores indicating a

person's personality, too can make individuals feel uncomfortable. The introspective process of measuring and quantifying ones internal states can feel intrusive and distressing. Particularly when certain traits, such as neuroticism, are deemed undesirable. Questionnaires conducted online, although often yielding a greater number of responses, don't allow for a debrief session between experimenter and participant (typical in lab experiments).

Recently, there has been a trend towards researchers in psychology attempting to predict people's personalities from their online social media accounts like Twitter (Quercia et al., 2011) and Facebook (Back et al., 2010). For Twitter research, this may include natural language processing from tweets, analysing topics from hashtags, or using activity logs (number of retweets, followers etc.). On Facebook, information on people's profiles, their 'likes', or their social network can be analysed. The guidelines around consent for this research are blurry (Kosinski and Behrend, 2017). For example, the individual partaking in the study may give permission for their Facebook profile to be analysed - but does this include permission to analyse other peoples comments on their profile, or their own comments on other peoples profiles? Furthermore, just because people are happy publicly posting things for others to see, does not necessarily mean they are happy for that data to be analysed. Schultze and Mason (2012) liken the logic to living in a small village. You may be aware that people might know information (sometimes private) about you, but that does not mean you want your neighbours discussing and analysing it.

Feeding algorithms personality information can improve targeted marketing (Matz and Netzer, 2017; Gladstone et al., 2019), making them rely less on sweeping generalisations of age and gender, and achieve greater personalisation. This can lead to greater customer satisfaction, as people are able to shop more efficiently and effectively. However, the flip-side of this is that target marketing can be viewed as deterministic, and improving it with personality data only amplifies this. Online, we experience a bubble effect, where we are exposed to information, individuals, and adverts which match our previously inferred views, likes, and beliefs. This creates a feedback loop, making it more likely we will continue to interact with this content in the future. Such exposure to only certain types of information can be dangerous as it limits our choices, and to some extent our free will.

As algorithms develop and the accuracy of predicting personality from digital footprints improves, worrying implications emerge. Firstly, there are some personality types which have the potential to be more easily exploited. In the Boots data, neuroticism was found to be the easiest trait to predict from shopping behaviour; and people high in neuroticism are probably also the most easily manipulated. For example, if a company wanted to market a weight loss supplement, their marketing would likely be most effective if targeted towards people who are higher in neuroticism (more likely to be insecure, guided by emotions, and have lower self-esteem). Additionally, if companies can predict neuroticism, or any trait to a high degree of accuracy, they won't need to measure every person's personality. This brings up issues of informed consent. At what accuracy does an algorithm have to get to, before we are unhappy about it 'guessing' things about us that we haven't given it permission to know?

Future legislation will have to deal with questions like this in order to protect vulnerable customers from being easily manipulated - akin to laws protecting problem gamblers from fixed-odds betting terminals (FOBTs) (Davies, 2016).

In summary, these are exciting yet volatile times for the use of big data in research. As data continues to rise in quantity and value, we are seeing companies, not governments, acquiring the greatest power. Looking forward, big data psychology research needs to continue to ensure all protocols, guidelines, and common sense is followed to ensure the discipline doesn't get tarnished, or confused, with the type of research viewed as unfavourable by the public (and that some companies are indeed doing). Furthermore, researchers should continue to communicate with the public and companies on the value of big data research and data donations - helping to ensure the potential of big data for public good is realised.

9.6 Conclusions on the Scientific Approach to Big Data Psychology

The previous section debates the advantages and disadvantages of big data, comparative to small data. In addition to the differences in the actual attributes of the data itself, big data differs from small data in terms of the scientific approaches applied to it: prediction versus explanation, and induction versus deduction (outlined in detail in Chapter 1). An aversion to an algorithmic modelling or inductive approach, for some in psychology, may be attributed in part to the current climate of the reproducibility crisis, and the high prevalence of *p*-hacking found in psychology (Open Science Collaboration, 2015). This context, may mean academics are more wary of methodologies resembling a “fishing exercise”. In contrast, an algorithmic modelling approach is common in computer science, often viewed as the only logical way to approach modelling in big data. The variable selection method of “chuck it all in and see what works” offers a great opportunity for knowledge discovery. While traditional psychologists may prefer simple models derived from the literature to prevent false discovery. The next section makes concluding comments on the utility of both approaches in big data psychology using the work presented in this thesis.

9.6.1 Prediction versus Explanation

Understanding human behaviour can encompass both explaining and predicting it. As discussed in Chapter 1, these two goals are not equivalent. Prediction involves finding the best model to forecast future observations. Whilst explanation requires identifying the causal processes which give rise to observed phenomena (Mahmoodi et al., 2017). However these goals are also not incompatible. For example, extracting variable importance from a prediction model or using explainable AI methods encompass both prediction and explanation. Yet often the two goals are not used in combination (Mahmoodi et al., 2017). As explanation and prediction require different scientific approaches, methodology, and statistical modelling,

they tend to be used to answer different questions (Shmueli et al., 2010). Whilst prediction models are used to answer applied questions in need of a practical solution, an explanation approach is used if the goal is greater knowledge.

Throughout the thesis, research explored achieving the goals of both prediction and explanation, as both are desirable for big data psychology: prediction prevents overfitting in large datasets (as well as identifying models which are operational), whilst an explanation approach can be used to develop psychological theories. This thesis has found that a useful way to combine the two, is to find the best (interpretable) prediction model, and extract the variable importance information from this. A suggested framework for doing this can be found in Figure 9.2.

The suggested approach involves optimising each model’s meta-parameters using cross fold validation and testing the models on out-of-sample data to compare their accuracy. The framework suggests using only relatively interpretable models like LRMs, DTs and RFs. Although other prediction models such as NNs exist, they are so complex that the mapping between the input and output is not retrievable (although work to try and remedy this continues apace). Although a metric of variable importance can be gained for NNs using permutations, the metric is highly influenced by multicollinearity (as surrogates are not typically harnessed, as they are in DTs and RFs), and it cannot be understood *how* a predictor is useful. For both NNs and LRMs, a variable could only be useful in a particular subgroup, or (particularly in the case of NNs) as a part of a complex network of interactions, but we wouldn’t know - thus a complete, actionable and useful explanation is hard to attain.

Once the optimal model solution is selected, here it is suggested that the model be fitted to 100% of the data to extract the variable importance. This ensures the most information (data) is available to the model (this is particularly important for decision trees as they are more unstable). Crucially, the model is prevented from overfitting at this stage by using the same meta-parameters which produced the best CV accuracy. This ensures that the variable importance information is retrieved from a generalisable model. At this point, different steps are recommended depending on the model selected in the prediction stage.

If the model selected is a regression model, and extracting statistically significant variables is a goal, then additional FDR should be controlled for before using the p -values as selection criteria. It is worth noting here, that it is possible the β coefficients may be inflated in this stage, as the model is being used for explanation, and thus fitted to the data. However, this is in line with the general practice in Psychological research when using regression to explain variance. Further, one of the meta-parameters which should be tuned in the Linear Regression Models’ (LRM) optimisation phase, is the regularisation parameter. A regularisation parameter, which has been selected to ensure the model generalises well, will help prevent inflation of β coefficients when the model is fitted to 100% of the data.

Although, prior to this point, when specifying the LRM, variable inflation factors (VIFs) would have been checked and multicollinearity addressed this way, an additional step (when using a LRM for explanation) is suggested to check for masking effects (where one variable

can prevent the effect of another variable being identified as important (refer to Chapter 7)). A quick check of this can be done by comparing the simple correlations with the regression coefficients. If, as in Section 6.3 of Chapter 6, the Pearson r values vary dramatically, or are even the opposite direction to the regression coefficients, post-hoc tests are recommended to investigate the potential masking/interaction effects. Such post-hoc tests will be able to provide a fuller, more detailed explanation.

Alternatively, if a Random Forest (RF) is selected in the prediction stage (it is highly unlikely a Decision Tree (DT) will produce more accurate out-of-sample predictions than a RF (Breiman, 2001)), it is possible that there exists sub-groups or non-linear relationships/interactions within the data. Although a RF is less interpretable than a LRM or a DT, variable importance rankings can still be attained from the model. The framework also proposes that in cases where 1) the DT's accuracy is not dramatically lower than the RF (within 5%, as a guideline), and 2) the variable importance rankings follow a similar pattern to the RF's (i.e. share the same most important predictors), then an assumption may be made - they are using a similar predictive mechanism (but the RF has smoother decision boundaries). In such cases, the DT structure can be used to reflect a simplified version of the RF's solution. This step is proposed, not for hypothesis testing, but as an exploratory further investigation into how the RF *may* be working and what sub groups, interactions, and non-linearities *may* be in the data.

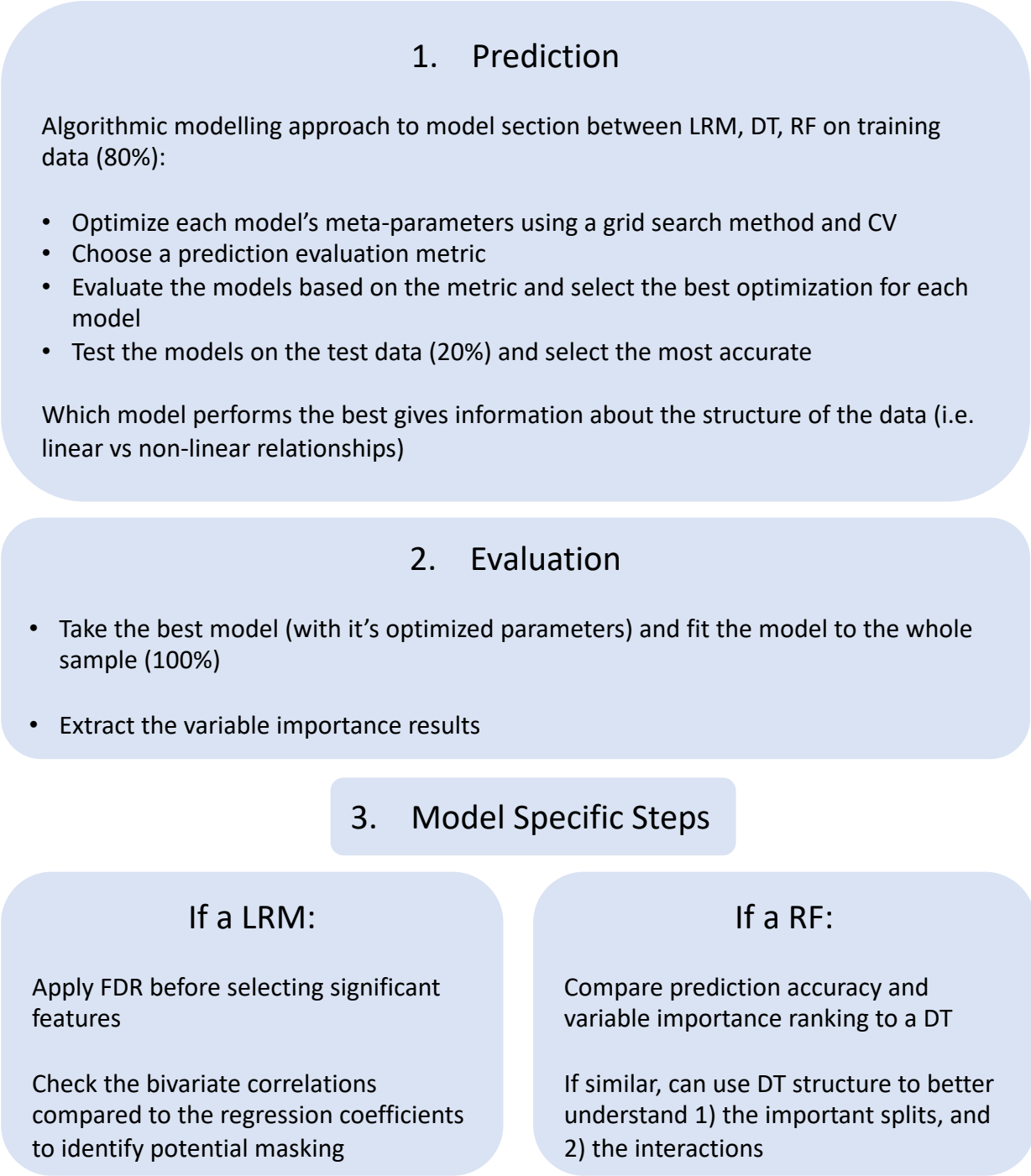


Figure 9.2: A proposed framework for using both prediction and explanation to understand predictive relationships in big data. LRM= linear regression model, DT= decision tree, RF= random forest, CV= cross validation. In step 1 of the framework (Prediction), an 80:20 split is suggested for splitting data into training:test sets, and CV for optimising meta-parameters, as these were the splits predominantly used in this thesis. However, other methods exist (for example 15% random sampling for model optimisation), and for different size and types of the data it is noted here that other data divisions may be more appropriate.

In summary, this research has presented a framework in which the otherwise separate goals, prediction and explanation, can be combined to retain accurate predictions whilst also allowing information to be extracted that can aid our understanding of the causal pathways. It is too much for one method to be able to give a complete picture of variable relationships as they are expected in the real world, as each model is constrained by its mechanism. However, different tools can chip away at uncovering the complexities of the

‘true’ picture. This picture involves understanding not just what predicts the DV best, but what variables may be being masked, mediating, or moderating other variables. Ultimately, these pathways are useful to identify in psychology so we can understand the causes of an outcome, and where best to intervene should we want to change that outcome.

What about Explainability in AI?

Above, a framework is laid out for combining both explanation and prediction in a way which is believed to be most relevant to big data psychology. This framework is suggested for logistic regressions, decision trees and random forests, and it is discussed why it is hard to use a black box algorithm, such as a Neural Network (NN), for explanation. However, limiting predictions to just using these simpler models may be at a cost to accuracy. Therefore, if such prediction models can somehow *also* be used for explanation, this has the potential to be preferable to the framework proposed above. Thus, this section outlines how those in the machine learning world are currently tackling the problem of explainability, and how this may (or may not) be useful in the realm of big data psychology.

Recently, the field of machine learning and artificial intelligence (AI) has put more focus on being able to explain and interpret models (Williams, 2018), and a new sub-field called explainable AI (XAI) has emerged as a result. This is in part due to a lack of trust by humans who interact with algorithms, and scandals of algorithm bias and discrimination emerging (Cossins, 2018). One pharmaceuticals company engineered an algorithm to optimise the yield of its factories, but its senior engineers didn’t trust it and refused to apply the proposed calibrations, resulting in the potential loss of \$10 to \$20 million (Williams, 2018). Yet, with more and more important decisions in the areas of employment, health, justice, policing, banking and insurance being made using algorithms, it is no wonder policymakers, businesses, and now the public are wanting to better understand how these decisions are being made.

Explainable AI is an exciting new area of development for achieving the goal of explanation and prediction, helping to “open the box” on how complex algorithms are making decisions. XAI methods include post-hoc approaches, which allow the model to be trained as normal, and then explainability is added on during the testing phase; and ante-hoc, which incorporates explainability into a model from the beginning. Ante-hoc methods include Reversed Time Attention Model (RETAIN) (Choi et al., 2016) for convolutional neural networks, and Bayesian Deep Learning (BDL) (Kendall and Gal, 2017) which capitalises on estimates of uncertainty. Whilst, post-hoc models like Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) can be applied to many different types of models after training to aid interpretation (and are available as open source packages). The advantage of the latter model-agnostic approach is that the initial deep model is not affected, so performance is not compromised for explanation. The black-box model is ran as usual, then either an interpretable model is learnt on the predictions of the black box model (Craven and Shavlik, 1996; Baehrens et al., 2010), or the inputs are perturbed (noise added) to seeing how the

black box model reacts (Kononenko et al., 2010; Krause et al., 2016), or both (Ribeiro et al., 2016).

The hope is that XAI will help to increase algorithmic fairness using the evaluations of the prediction reasons to identify bias, and then have humans (when necessary) override/correct such parameters manually, in a hybrid approach. The progress being made in XAI is exciting not just for the transparency of algorithms, but also their improvement. If we understand how something works better, we can better understand how to improve it.

With all this potential of XAI in the field of machine learning, can it be used in big data psychology to understand human behaviour? The answer to this depends on what qualifies as a satisfying explanation. XAI understands how a deep-learning model works, but that is mostly what it is limited to. In psychology, most want to understand the relationships between an IV (or a set of IVs) and a DV, as reflected in the world. This often requires theories which incorporate the findings from a model into our existing world knowledge/beliefs retrieved from other models. A deep-learning model can make an accurate prediction, but the explanation derived from the model can be causally impossible (Mahmoodi et al., 2017). Take the example of predicting intelligence using likes on Facebook. The algorithm may be able to give an accurate prediction, but if it explains the prediction by telling us that liking “curly fries” is the most important feature for predicting IQ (Kosinski et al., 2013), this is not useful in the advancement of understanding the construct of intelligence.

A deep-learning model is not a direct mapping of what is occurring in the real world, the model will learn to take shortcuts. Lets take the example of age being used to predict impulsivity. The strength of age could mask the effect of other variables which co-occur with age. In reality, the variable labelled ‘age’ represents, not just the number of years in existence, but experience, life events, and knowledge (to name just a few of the things that co-vary with age). Psychology is more interested in answering questions such as ‘what are the underlying variables or constructs which co-vary with age, and lead to the reduction in impulsivity later in life?’. For this reason, post-hoc analyses to identify masking is recommend in the framework in Figure 9.2, to try to get as full a picture as possible on the causal relationships as they exist in the world.

In summary, XAI does not offer an instant solution to the goal of trying to explain and predict behaviour. In reality, there will always likely be some compromise between explanation and prediction. It is probable that the world is exceedingly more complex than we as humans can perceive it. Thus we require models which go beyond what we can understand to model it. However, a complex model is most likely going to require a complex theory to explain it. And for theories on human behaviour to be useful (for us to use them to explain and predict future behaviour), they need to apply a substantial simplification to complex behaviours in the real world. The prevailing current solution to the type of explanation required for psychologists to further theory is to use simpler, more interpretable models, such as regressions and decisions trees. Thus, the framework suggested in Figure 9.2 aims to supplement what is already being used in psychology, and promote the use of

prediction to prevent overfitting, in corporation with existing explanation based approaches.

9.6.2 Induction versus Deduction

As with the trade-off between prediction and explanation, induction versus deduction is another cultural clash which is apparent when the fields of psychology and machine learning come together. As outlined in Chapter 1, induction and deduction are two opposing approaches to scientific knowledge discovery. Deduction is the common approach in psychology, where a researcher thinks of a theory/logic they want to test, and goes out and collects data with the goal of rejecting the null hypothesis (adding evidence to their theory). This is a top-down theory/logic driven approach. Induction is the opposite, and is more common in computer science. An inductive approach uses the data or observations available to induce laws or a theory as to how the observations are occurring.

The choice between using an inductive or deductive approach is intrinsically linked to the small data versus big data debate. Induction requires data to exist, and works better on large and representative datasets, where effects can be detected reliably (Mahmoodi et al., 2017). With large datasets available for research becoming more common, some believe that knowledge driven methods will be replaced by big data and induction (Lazer et al., 2014), even claiming that big data makes deduction outdated (Anderson, 2008).

Throughout my research, I have noted that journal reviewers have had particular issue with an approach which “ignores theory”. *A priori* theories in psychology are undoubtedly useful as (among other things) they provide findings with a meaningful context, help to summarise findings down to a few core ideas, and act as a springboard for generating new hypotheses to test. Yet being constrained by a theory *only* approach isn’t always helpful. One of the greatest breakthroughs in personality psychology was through a bottom-up inductive methodology - Goldberg’s lexical structure of the Big Five personality traits (Goldberg, 1990). Unsurprisingly, despite it being the most widely used model of personality, its main criticism is the lack of theory it was derived from. Yet, this is the perfect example of theory not being needed to render something exceedingly useful. The five factor structure has subsequently led to many theoretical findings, showing that utility in itself is a finding, and one that can feed into future theory generation. Many believe the goal of utility, and consequently prediction, should be driving more research in psychology. As Goldberg’s five factor structure shows, prediction is useful in its own right, whilst there is little point in having something that makes sense theoretically, but is unusable. In a sense, Goldberg (1990)’s method in itself is a theory - a theory that mining human language can reveal some structure about our internal intangible qualities. Thus, a method can be applied, and if it works, the theoretical implications as to why it works can always be worked out later.

As with the debate on prediction versus explanation, this research has proposed an approach to big data psychology where researchers can utilise both inductive and deductive methods. Figure 9.3 illustrates how induction can supplement the deductive approach pre-

dominately used in psychology, when the two are combined into an iterative cycle. On the left of the diagram (depicting deduction first, followed by induction) is the methodology used in Chapter 4 (predicting SWB).

In Chapter 4, hypotheses were first made and tested on the relationship between PA and SWB based on previous literature (primarily Matz et al. (2016)’s study). Following this, other machine learning models were tested to check that the model selected (a linear regression) was the best model. Despite the LRM performing similarly to the RF using CV, PA^N was ranked as much more important in the RF model. This lead to post-hoc analyses finding that the relationship between PA^N and SWB significantly fitted a non-linear polynomial function. Thus, the second inductive stage (as illustrated on the left hand side of the diagram in Figure 9.3, deduction→induction) checked, and called into question, the assumption that PA would linearly predict SWB.

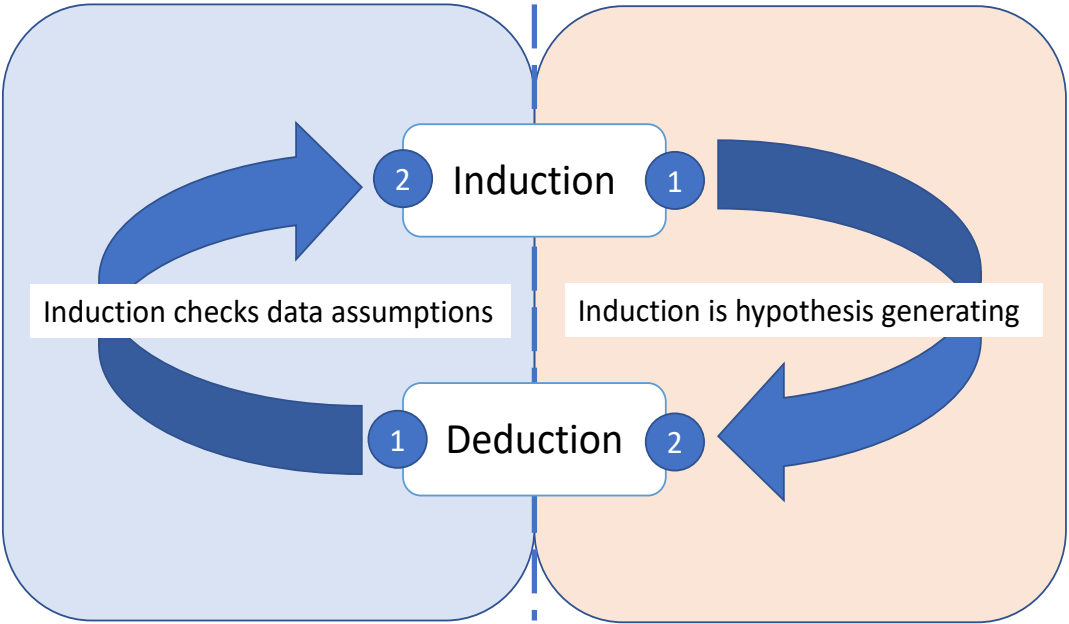


Figure 9.3: A proposed approach for using both induction and deduction in big data psychology, illustrating how induction can supplement the deductive approach currently being predominantly used in psychology.

The right side of the diagram, depicted in Figure 9.3, represents how an inductive approach can supplement deduction, by being hypothesis generating. This could simply involve hypothesising that (as in the above example) a variable, such as PA^N , has a non-linear relationship with the DV (SWB). Or, it could be more complex, as in Chapter 7, where after conducting multiple analyses, it was hypothesised that impulsivity (measured using Behavioural Activation System (BAS)) *mediated* the effect of age on plastic bag buying. This was then explicitly tested using a Structural Equation Model (SEM). These examples, (for which there are more in this thesis) highlight the need to find the right models for the right hypotheses, and how analysing the differences in the outputs between multiple analyses (whilst being aware of the different models capabilities and limitations) can be particularly insightful, and ultimately lead to the generation of new hypotheses and theories.

Ultimately, “Science is a never-ending, always changing process” (Rothchild, 2006) and

scientists should use the approach which is most apposite to the nature of the data at hand. In some cases, this could require proponents of a theory-driven, deductive-only approach in psychology to be open-minded to the idea of induction, given big data is now more prevalent in research. Researchers should be encouraged to be flexible, and creative with their application of methods for uncovering knowledge. As Emmert Wolf said, “A man is only as good as his tools” - psychologists should be reassured and encouraged that induction is a legitimate and useful tool to have under their belt in the era of big data.

To conclude, psychologists should not limit themselves to either induction or deduction. Just as scientists don’t want knowledge discovery to be restricted, the tools to undergo it should not be either. Thus, both inductive and deductive approaches can be useful in big data psychology, and the choice doesn’t have to be between them - with enough data for exploration, whilst also keeping some back for testing, the two can be combined into an iterative cycle .

9.7 Concluding Remarks

In Chapter 2, this thesis identified the following gaps in the big data psychology literature. Firstly, that little psychology research in big data is conducted by psychologists for psychology journals, and such research is rarely occurring outside of a consumer/marketing domain. Furthermore, the research is focused on exploiting resources, such as personality data to improve prediction models, and less about using these resources to interrogate the constructs themselves and their relationship with behaviour in big data (Bleidorn et al., 2017). Secondly, there is a lack of engagement in debates about the nuances as to when and why big/small data approaches and analyses are appropriate. Methods journals in psychology have published tutorials and examples of how psychologists can extract and process big data to analyse. But it is rarely discussed *why* methods (such as cross-validation) are useful, or the fundamental ways in which small data and big data vary, and subsequently the different approaches required. Finally, there is a lack of investigation into the psychological drivers of behaviour on an *individual-level*. This is likely due to data privacy and availability restraints.

As a result, this thesis exploited the unique opportunity afforded by Boots to link real world transaction data with questionnaire responses measuring personality and other psychological phenomena. This resource enabled the study of individual processes *en masse* - a unique and exciting opportunity for a psychologist. To address the lack of education about the differences between big and small data approaches, this thesis attempted to engage psychologists in the advantages and disadvantages of big data, and to critically discuss the methodologies used in big data research that are different from those traditionally used in psychology (namely out-of-sample prediction, machine learning models, cross-validation, an algorithmic modelling approach, model evaluation and selection based on performance, and induction). Finally, in an attempt to contribute towards bridging the gap between psy-

chologists and big data research (as outlined in Chapter 2), this thesis aimed to take the reader on a journey into the areas of big data and machine learning identified as being most relevant and useful to psychologists (from a tradition schooling), from the perspective of a psychologist, who has made the very same journey. Such a perspective may be an important part of engaging the field of psychology more, so psychologists can “be proactive and take advantage of a real opportunity in front of them” (Harlow and Oswald, 2016), by actively contributing to big data psychology.

But why does it matter whether psychologists play a part in big data psychology research? Here, it is argued that psychologists conceivably have much to contribute, both to big data psychology, and further to the field of machine learning. Firstly, psychologists, who have a deeper knowledge of the nuanced literature on the psychological phenomena that predict behaviour, can help incorporate such knowledge into models via model *inputs*. We are already seeing the success of this in consumer psychology, where personality data (mainly the Big Five) is being used to improve prediction models (i.e. Tkalcic et al. (2009)). Yet, the Big Five personality framework is only a slice of what psychology, a discipline which specialises in theories of human behaviour, has to offer. Literatures on “nudges” (Thaler and Sunstein, 2009), cognitive biases (Tversky and Kahneman, 1974), and behavioural reinforcement (Corr, 2004) (among others) could be exploited to try to improve models of human behaviour. Secondly, the drive in psychology to *explain* the world using models, highlights the deficiencies in the ability to interpret models in machine learning. Particularly variable importance measures under conditions of high multicollinearity (when masking can occur). Using their in-depth knowledge of psychological theory to interpret *outputs* of models of human behaviour, leading to a greater understanding, psychologists could help the development of better (more accurate or more parsimonious) prediction models. Finally, from improved explainability of the outputs could be drawn *theoretical* contributions, which could in turn contribute back into model development/specification. In conclusion, as well as psychologists potentially using big data to provide measures of real world behaviour, for large sample sizes, and prediction methods to prevent overfitting, researchers in machine learning can also use psychology to improve the predictions and explainability of their behavioural prediction models.

Bibliography

- Aaker, J. L. (1997). Dimensions of brand personality. *Journal of marketing research*, pages 347–356.
- Abdulla, N., Ahmed, Z., and Gazzali, M. (2015). Measure customer behaviour using c4.5 decision tree map reduce implementation in big data analytics and data visualization. *International Journal for Innovative Research in Science & Technology*, pages 228–235.
- Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, 97(1):129.
- Abrahams, B. (1997). It’s all in the mind. *Marketing*, 27(1):31–33.
- Abrahamse, W., Steg, L., Gifford, R., and Vlek, C. (2009). Factors influencing car use for commuting and the intention to reduce it: A question of self-interest or morality? *Transportation Research Part F: Traffic Psychology and Behaviour*, 12(4):317–324.
- Ackerman, P. L. and Heggestad, E. D. (1997). Intelligence, personality, and interests: evidence for overlapping traits. *Psychological bulletin*, 121(2):219.
- Adams, Z. W., Kaiser, A. J., Lynam, D. R., Charnigo, R. J., and Milich, R. (2012). Drinking motives as mediators of the impulsivity-substance use relation: Pathways for negative urgency, lack of premeditation, and sensation seeking. *Addictive behaviors*, 37(7):848–855.
- Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, (6):734–749.
- Adrian, J. and Daniel, R. (1976). Impact of socioeconomic factors on consumption of selected food nutrients in the united states. *American Journal of Agricultural Economics*, 58(1):31–38.
- Albinsson, P. A., Wolf, M., and Kopf, D. A. (2010). Anti-consumption in east germany: consumer resistance to hyperconsumption. *Journal of Consumer Behaviour*, 9(6):412–425.
- Allison, P. (2012). When can you safely ignore multicollinearity? <https://statisticalhorizons.com/multicollinearity>.

- Anastasio, M. and Nix, J. (2016). Plastic bag levy in ireland. <https://www.nidirect.gov.uk/articles/carrier-bag-levy>.
- Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07.
- Ando, K., Ohnuma, S., Blöbaum, A., Matthies, E., and Sugiura, J. (2010). Determinants of individual and collective pro-environmental behaviors: Comparing germany and japan. *Journal of environmental information science*, 38(5):21–32.
- Angst, J. and Dobler-Mikola, A. (1985). The zurich study. v. anxiety and phobia in young adults. *European archives of psychiatry and neurological sciences*, 235(3):171–178.
- Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical analysis*, 27(2):93–115.
- Anseau, M., Dierick, M., Buntinx, F., Cnockaert, P., De Smedt, J., Van Den Haute, M., and Vander Mijnsbrugge, D. (2004). High prevalence of mental disorders in primary care. *Journal of affective disorders*, 78(1):49–55.
- Arbuthnot, J. (1977). The roles of attitudinal and personality variables in the prediction of environmental behavior and knowledge. *Environment and behavior*, 9(2):217–232.
- Arcury, T. A. and Christianson, E. H. (1993). Rural-urban differences in environmental knowledge and actions. *The Journal of Environmental Education*, 25(1):19–25.
- Ariely, D. and Wertenbroch, K. (2002). Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological science*, 13(3):219–224.
- Armen, A. P. and Tsamardinos, I. (2011). A unified approach to estimation and control of the false discovery rate in bayesian network skeleton identification. In *ESANN*.
- Arnett, J. J. (2008). The neglected 95%: why american psychology needs to become less american. *American Psychologist*, 63(7):602.
- Asendorpf, J. B. (1992). Continuity and stability of personality traits and personality patterns.
- Ashton, M. C., Lee, K., and Goldberg, L. R. (2004). A hierarchical analysis of 1,710 english personality-descriptive adjectives. *Journal of Personality and Social Psychology*, 87(5):707.
- Atkinson, E. J. and Therneau, T. M. (2000). An introduction to recursive partitioning using the rpart routines. *Rochester: Mayo Foundation*.
- Australia, C. U. (2015). Report on actions to reduce circulation of single-use plastic bags around the world: August 2015 (retrieved from australia).
- Avis, M., Forbes, S., and Ferguson, S. (2014). The brand personality of rocks: A critical evaluation of a brand personality scale. *Marketing Theory*, 14(4):451–475.

- Axon, R. N., Bradford, W. D., and Egan, B. M. (2009). The role of individual time preferences in health behaviors among hypertensive adults: a pilot study. *Journal of the American Society of Hypertension*, 3(1):35–41.
- Axon, S. (2017). “keeping the ball rolling”: Addressing the enablers of, and barriers to, sustainable lifestyles. *Journal of Environmental Psychology*, 52:11–25.
- Ayers, J. W., Althouse, B. M., Allem, J.-P., Rosenquist, J. N., and Ford, D. E. (2013). Seasonality in seeking mental health information on google. *American journal of preventive medicine*, 44(5):520–525.
- Baca-Garcia, E., Perez-Rodriguez, M. M., Saiz-Gonzalez, D., Basurte-Villamor, I., Saiz-Ruiz, J., Leiva-Murillo, J. M., de Prado-Cumplido, M., Santiago-Mozos, R., Artés-Rodríguez, A., and De Leon, J. (2007). Variables associated with familial suicide attempts in a sample of suicide attempters. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 31(6):1312–1316.
- Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., and Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological science*, 21(3):372–374.
- Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and MÄžller, K.-R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831.
- Bagwell, L. S. and Bernheim, B. D. (1996). Veblen effects in a theory of conspicuous consumption. *The American Economic Review*, pages 349–373.
- Balderjahn, I. (1988). Personality variables and environmental attitudes as predictors of ecologically responsible consumption patterns. *Journal of business Research*, 17(1):51–56.
- Baldwin, C. (2018). Why the boots advantage card works so well. <https://blog.sodexoengage.com/consumer-loyalty/why-the-boots-advantage-card-works-so-well>.
- Bamberg, S. and Möser, G. (2007). Twenty years after hines, hungerford, and tomara: A new meta-analysis of psycho-social determinants of pro-environmental behaviour. *Journal of environmental psychology*, 27(1):14–25.
- Banerjee, A. V. and Duflo, E. (2007). The economic lives of the poor. *Journal of economic perspectives*, 21(1):141–168.
- Baumeister, R. F. (2002). Yielding to temptation: Self-control failure, impulsive purchasing, and consumer behavior. *Journal of consumer Research*, 28(4):670–676.
- Baumeister, R. F., Bratslavsky, E., Muraven, M., and Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of personality and social psychology*, 74(5):1252.

- Bearden, W. O., Money, R. B., and Nevins, J. L. (2006). A measure of long-term orientation: Development and validation. *Journal of the Academy of Marketing Science*, 34(3):456–467.
- Bearden, W. O. and Netemeyer, R. G. (1999). *Handbook of marketing scales: Multi-item measures for marketing and consumer behavior research*. Sage.
- Beatty, S. E. and Ferrell, M. E. (1998). Impulse buying: Modeling its precursors. *Journal of retailing*, 74(2):169–191.
- Belk, R. W. (1984). Three scales to measure constructs related to materialism: Reliability, validity, and relationships to measures of happiness. *ACR North American Advances*.
- Bellenger, D. N., Robertson, D. H., and Hirschman, E. C. (1978). Impulse buying varies by product. *Journal of Advertising Research*, 18(6):15–18.
- Bellizzi, J. A. and Bristol, T. (2004). An assessment of supermarket loyalty cards in one major us market. *Journal of Consumer Marketing*, 21(2):144–154.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.
- Bhatnagar, A., Misra, S., and Rao, H. R. (2000). On risk, convenience, and internet shopping behavior. *Communications of the ACM*, 43(11):98–105.
- Bickel, P. J., Ritov, Y., Stoker, T. M., et al. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *The Annals of Statistics*, 34(2):721–741.
- Bishai, D. (2001). Lifecycle changes in the rate of time preference: testing the theory of endogenous preferences and its relevance to adolescent substance abuse. *Economic Analysis of Substance Abuse*, pages 155–77.
- Bishop, C. M. (2006). Pattern recognition and machine learning (information science and statistics) springer-verlag new york. *Inc. Secaucus, NJ, USA*.
- Black, D. W. (2001). Compulsive buying disorder. *CNS drugs*, 15(1):17–27.
- Black, D. W., Repertinger, S., Gaffney, G. R., and Gabel, J. (1998). Family history and psychiatric comorbidity in persons with compulsive buying: preliminary findings. *American Journal of Psychiatry*, 155(7):960–963.
- Blakeman, S. and Mareschal, D. (2020). A complementary learning systems approach to temporal difference learning. *Neural Networks*, 122:218–230.
- Blaylock, J., Smallwood, D., Kassel, K., Variyam, J., and Aldrich, L. (1999). Economics, food choices, and nutrition. *Food Policy*, 24(2-3):269–286.

- Bleidorn, W., Hopwood, C. J., Ackerman, R. A., Witt, E. A., Kandler, C., Riemann, R., Samuel, D. B., and Donnellan, M. B. (2019). The healthy personality from a basic trait perspective. *Journal of personality and social psychology*.
- Bleidorn, W., Hopwood, C. J., and Wright, A. G. (2017). Using big data to advance personality theory. *Current Opinion in Behavioral Sciences*, 18:79–82.
- Block, B. (2013). China reports 66-percent drop in plastic bag use. *Eye on Earth, Worldwatch Institute*. Retrieved November, 29.
- Boneva, B., Frieze, I. H., Ferligoj, A., Jarošová, E., Pauknerová, D., and Orgocka, A. (1998). Achievement, power, and affiliation motives as clues to (e) migration desires: A four-countries comparison. *European Psychologist*, 3(4):247–254.
- Boots (2019). Boots in numbers. <https://www.boots-uk.com/about-boots-uk/about-boots/boots-in-numbers/>.
- Borden, R. J. and Francis, J. L. (1978). Who cares about ecology? personality and sex differences in environmental concern 1. *Journal of Personality*, 46(1):190–203.
- Borsboom, D., Kievit, R. A., Cervone, D., and Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In *Dynamic process methodology in the social and developmental sciences*, pages 67–97. Springer.
- Bove, L. L., Nagpal, A., and Dorsett, A. D. S. (2009). Exploring the determinants of the frugal shopper. *Journal of Retailing and Consumer Services*, 16(4):291–297.
- Bower, G. H. (1981). Mood and memory. *American psychologist*, 36(2):129.
- Boyce, C. J., Daly, M., Hounkpatin, H. O., and Wood, A. M. (2017). Money may buy happiness, but often so little that it doesn’t matter. *Psychological science*, 28(4):544–546.
- Bradford, D. and McAlvanah, P. (2014). Time preferences, health behaviors, and energy consumption.
- Bradford, W. D. (2010). The association between individual time preferences and health maintenance habits. *Medical Decision Making*, 30(1):99–112.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.
- Breiman, L., Friedman, H., Olshen, A., and Stone, C. (1984). Classification and regression trees. monterey ca: Wadsworth & brooks/cole advanced books & software.
- Brown, T. A. and Barlow, D. H. (2009). A proposal for a dimensional classification system based on the shared features of the dsm-iv anxiety and mood disorders: Implications for assessment and treatment. *Psychological assessment*, 21(3):256.

- Bruce, S. E., Yonkers, K. A., Otto, M. W., Eisen, J. L., Weisberg, R. B., Pagano, M., Shea, M. T., and Keller, M. B. (2005). Influence of psychiatric comorbidity on recovery and recurrence in generalized anxiety disorder, social phobia, and panic disorder: a 12-year prospective study. *American Journal of psychiatry*, 162(6):1179–1187.
- Bruckman, A. (2002). Studying the amateur artist: A perspective on disguising data collected in human subjects research on the internet. *Ethics and Information Technology*, 4(3):217–231.
- Buchanan, E., Aycock, J., Dexter, S., Dittrich, D., and Hvizdak, E. (2011). Computer science security research and human subjects: Emerging considerations for research ethics boards. *Journal of Empirical Research on Human Research Ethics*, 6(2):71–83.
- Burger, J. M. and Caldwell, D. F. (2000). Personality, social activities, job-search behavior and interview success: Distinguishing between panas trait positive affect and neo extraversion. *Motivation and Emotion*, 24(1):51–62.
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365.
- Byrnes, J. P., Miller, D. C., and Schafer, W. D. (1999). Gender differences in risk taking: a meta-analysis. *Psychological bulletin*, 125(3):367.
- Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics versus machine learning. *Nature methods*, 15(4):233.
- Calkins, A. W., Heaton, B. A., Capozzoli, M. C., and Otto, M. W. (2013). Psychosocial predictors of sleep dysfunction: the role of anxiety sensitivity, dysfunctional beliefs, and neuroticism. *Behavioral Sleep Medicine*, 11(2):133–143.
- Carter, T. J. and Gilovich, T. (2010). The relative relativity of material and experiential purchases. *Journal of personality and social psychology*, 98(1):146.
- Carver, C. S., Sutton, S. K., and Scheier, M. F. (2000). Action, emotion, and personality: Emerging conceptual integration. *Personality and social psychology bulletin*, 26(6):741–751.
- Carver, C. S. and White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the bis/bas scales. *Journal of personality and social psychology*, 67(2):319.
- Casey, B. (2015). Beyond simple models of self-control to circuit-based accounts of adolescent behavior. *Annual review of psychology*, 66:295–319.

- Chabris, C. F., Laibson, D., Morris, C. L., Schuldt, J. P., and Taubinsky, D. (2008). Individual laboratory-measured discount rates predict field behavior. *Journal of risk and uncertainty*, 37(2-3):237.
- Chao, Y.-L. and Lam, S.-P. (2011). Measuring responsible environmental behavior: Self-reported and other-reported measures and their differences in testing a behavioral model. *Environment and Behavior*, 43(1):53–71.
- Chapman, G. B. and Elstein, A. S. (1995). Valuing the future: Temporal discounting of health and money. *Medical decision making*, 15(4):373–386.
- Chaudhuri, A. and Holbrook, M. B. (2001). The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty. *Journal of marketing*, 65(2):81–93.
- Chen, E. E. and Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, 21(4):458.
- Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. (2016). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512.
- Chu, K. K. A. (2015). Two-year worth of smart card transaction data—extracting longitudinal observations for the understanding of travel behaviour. *Transportation Research Procedia*, 11:365–380.
- Clapp, J. and Swanston, L. (2009). Doing away with plastic shopping bags: international patterns of norm emergence and policy implementation. *Environmental politics*, 18(3):315–332.
- Clark, A. E., Frijters, P., and Shields, M. A. (2008). Relative income, happiness, and utility: An explanation for the easterlin paradox and other puzzles. *Journal of Economic literature*, 46(1):95–144.
- Clarke, J. (2014). Plastic bags - the irish experience. *Paper presented at the dialogue event “disposable bag”, brussels (24/02/2014)*.
- Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.
- Clifford, L. (2008). Big data: How do your data grow? *Nature*, 455(7209):28–29.
- Cobb-Clark, D. A. and Schurer, S. (2012). The stability of big-five personality traits. *Economics Letters*, 115(1):11–15.
- Coley, A. and Burgess, B. (2003). Gender differences in cognitive and affective impulse buying. *Journal of Fashion Marketing and Management: An International Journal*, 7(3):282–295.

- Convery, F., McDonnell, S., and Ferreira, S. (2007). The most popular tax in europe? lessons from the irish plastic bags levy. *Environmental and resource economics*, 38(1):1–11.
- Corr, P. J. (2004). Reinforcement sensitivity theory and personality. *Neuroscience & Biobehavioral Reviews*, 28(3):317–332.
- Corr, P. J., Pickering, A. D., and Gray, J. A. (1995). Personality and reinforcement in associative and instrumental learning. *Personality and individual Differences*, 19(1):47–71.
- Corral-Verdugo, V., Fraijo-Sing, B., and Pinheiro, J. Q. (2006). Sustainable behavior and time perspective: Present, past, and future orientations and their relationship with water conservation behavior. *Interamerican Journal of Psychology*, 40(2):139–147.
- Cortina, J. M. and Landis, R. S. (2009). When small effect sizes tell a big story, and when large effect sizes don’t. *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*, pages 287–308.
- Cossins, D. (2018). Discriminating algorithms: 5 times ai showed prejudice. <https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/#ixzz634tKf0uX>.
- Costa, P. T. and MacCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources, Incorporated.
- Costa, P. T. and McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: happy and unhappy people. *Journal of personality and social psychology*, 38(4):668.
- Costa, P. T. and McCrae, R. R. (1985). The neo personality inventory.
- Costa Jr, P. T. and McCrae, R. R. (2008). The revised neo personality inventory (neo-pi-r).
- Costa Jr, P. T., McCrae, R. R., and Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: A revision of the neo personality inventory. *Personality and individual Differences*, 12(9):887–898.
- Costa Jr, P. T., Terracciano, A., and McCrae, R. R. (2001). Gender differences in personality traits across cultures: robust and surprising findings. *Journal of personality and social psychology*, 81(2):322.
- Craven, M. and Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*, pages 24–30.
- Cuartas Arias, J. M. (2017). Big data for use in psychological research. *International Journal of Psychological Research*, 10(1):6–7.

- Culotta, A. and Cutler, J. (2016). Mining brand perceptions from twitter social networks. *Marketing science*, 35(3):343–362.
- Cyders, M. A. and Coskunpinar, A. (2012). The relationship between self-report and lab task conceptualizations of impulsivity. *Journal of Research in Personality*, 46(1):121–124.
- Cyders, M. A., Smith, G. T., Spillane, N. S., Fischer, S., Annus, A. M., and Peterson, C. (2007). Integration of impulsivity and positive mood to predict risky behavior: development and validation of a measure of positive urgency. *Psychological assessment*, 19(1):107.
- DAERA (2018). Northern ireland carrier bag levy statistics. <https://www.daera-ni.gov.uk/articles/northern-ireland-carrier-bag-levy-statistics>.
- Dahl, F. A., Grotle, M., Benth, J. Š., and Natvig, B. (2008). Data splitting as a counter-measure against hypothesis fishing: with a case study of predictors for low back pain. *European journal of epidemiology*, 23(4):237–242.
- Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318. ACM.
- d’Astous, A. (1990). An inquiry into the compulsive side of “normal” consumers. *Journal of Consumer Policy*, 13(1):15–31.
- Datta, A., Tschantz, M. C., and Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies*, 2015(1):92–112.
- Davenport, T. H., Harris, J. G., and Kohli, A. K. (2001). How do they know their customers so well? *MIT Sloan Management Review*, 42(2):63.
- Davidson, D. J. and Freudenburg, W. R. (1996). Gender and environmental risk concerns: A review and analysis of available research. *Environment and behavior*, 28(3):302–339.
- Davidson, R. J. et al. (2001). Toward a biology of personality and emotion. *ANNALS-NEW YORK ACADEMY OF SCIENCES*, 935:191–207.
- Davies, R. (2016). Cut fixed-odds betting stakes to £2, mps urge.
- Day, G. S. (1976). A two-dimensional concept of brand loyalty. In *Mathematical models in marketing*, pages 89–89. Springer.
- de Chavez, A. C., Backett-Milburn, K., Parry, O., and Platt, S. (2005). Understanding and researching wellbeing: Its usage in different disciplines and potential for health research and health promotion. *Health Education Journal*, 64(1):70–87.
- De Dominicis, S., Sokoloski, R., Jaeger, C. M., and Schultz, P. W. (2019). Making the smart meter social promotes long-term energy conservation. *Palgrave Communications*, 5(1):51.

- Deeming, C. (2013). Addressing the social determinants of subjective wellbeing: the latest challenge for social policy. *Journal of social policy*, 42(3):541–565.
- DEFRA (2018a). Plastic carrier bags: Gove sets out new measures to extend charge. <https://www.gov.uk/government/news/plastic-carrier-bags-gove-sets-out-new-measures-to-extend-charge>.
- DEFRA (2018b). Single-use plastic carrier bags charge: data in england for 2016 to 2017. <https://www.gov.uk/government/publications/carrier-bag-charge-summary-of-data-in-england/single-use-plastic-carrier-bags-charge-data-in-england-for-2016-to-2017>.
- DellaVigna, S. and Malmendier, U. (2006). Paying not to go to the gym. *american economic Review*, 96(3):694–719.
- Deng, H. (2018). Why random forests outperform decision trees: Two intuitive reasons illustrated. <https://towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0b5>.
- DeSarbo, W. S. and Edwards, E. A. (1996). Typologies of compulsive buying behavior: A constrained clusterwise regression approach. *Journal of Consumer Psychology*, 5(3):231–262.
- DeYoung, C. G., Peterson, J. B., and Higgins, D. M. (2005). Sources of openness/intellect: Cognitive and neuropsychological correlates of the fifth factor of personality. *Journal of personality*, 73(4):825–858.
- DeYoung, C. G., Shamosh, N. A., Green, A. E., Braver, T. S., and Gray, J. R. (2009). Intellect as distinct from openness: differences revealed by fmri of working memory. *Journal of personality and social psychology*, 97(5):883.
- Deza, M. M. and Deza, E. (2009). Encyclopedia of distances. In *Encyclopedia of Distances*, pages 1–583. Springer.
- Di Tella, R. and MacCulloch, R. (2008). Gross national happiness as an answer to the easterlin paradox? *Journal of Development Economics*, 86(1):22–42.
- Diamantopoulos, A., Schlegelmilch, B. B., Sinkovics, R. R., and Bohlen, G. M. (2003). Can socio-demographics still play a role in profiling green consumers? a review of the evidence and an empirical investigation. *Journal of Business research*, 56(6):465–480.
- Dickman, S. J. (1990). Functional and dysfunctional impulsivity: personality and cognitive correlates. *Journal of personality and social psychology*, 58(1):95.
- Diener, E., Lucas, R. E., Oishi, S., et al. (2002). Subjective well-being: The science of happiness and life satisfaction. *Handbook of positive psychology*, 2:63–73.

- Diener, E., Oishi, S., and Lucas, R. E. (2003). Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life. *Annual review of psychology*, 54(1):403–425.
- Diener, E. and Seligman, M. E. (2004). Beyond money: Toward an economy of well-being. *Psychological science in the public interest*, 5(1):1–31.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.
- Dittmar, H., Beattie, J., and Friese, S. (1995). Gender identity and material symbols: Objects and decision considerations in impulse purchases. *Journal of economic psychology*, 16(3):491–511.
- Dittmar, H., Bond, R., Hurst, M., and Kasser, T. (2014). The relationship between materialism and personal well-being: A meta-analysis. *Journal of personality and social psychology*, 107(5):879.
- Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.
- Duijsens, I. J. and Diekstra, R. W. (1996). Dsm-iii-r and icd-10 personality disorders and their relationship with the big five dimensions of personality. *Personality and individual differences*, 21(1):119–133.
- Duncan, G. (2005). What do we mean by” happiness”? the relevance of subjective wellbeing to social policy. *Social Policy Journal of New Zealand*, 25:16.
- Dunlap, R. E. and Van Liere, K. D. (1978). The “new environmental paradigm”. *The journal of environmental education*, 9(4):10–19.
- Dunlap, R. E., Van Liere, K. D., Mertig, A. G., and Jones, R. E. (2000). New trends in measuring environmental attitudes: measuring endorsement of the new ecological paradigm: a revised nep scale. *Journal of social issues*, 56(3):425–442.
- Dunn, E. W., Aknin, L. B., and Norton, M. I. (2014). Prosocial spending and happiness: Using money to benefit others pays off. *Current Directions in Psychological Science*, 23(1):41–47.
- Dunn, E. W., Gilbert, D. T., and Wilson, T. D. (2011). If money doesn’t make you happy, then you probably aren’t spending it right. *Journal of Consumer Psychology*, 21(2):115–125.
- Dunson, D. B. (2018). Statistics in the big data era: Failures of the machine. *Statistics & Probability Letters*, 136:4–9.
- Dzogang, F., Goulding, J., Lightman, S., and Cristianini, N. (2017). Seasonal variation in collective mood via twitter content and medical purchases. In *International Symposium on Intelligent Data Analysis*, pages 63–74. Springer.

- Edwards, C. and Fry, J. M. (2011). Life cycle assessment of supermarket carrier bags. *Environment Agency, Horizon House, Deanery Road, Bristol, BS1 5AH*.
- Egol, M., Clyde, A., Rangan, K., and Sanderson, R. (2010). The new consumer frugality: adapting to the enduring shift in us consumer spending and behavior. *Booz & Company*, 24.
- Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2017). Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*.
- Fang, H. and Silverman, D. (2009). Time-inconsistency and welfare program participation: Evidence from the nlsy. *International Economic Review*, 50(4):1043–1077.
- Faraway, J. J. and Augustin, N. H. (2018). When small data beats big data. *Statistics & Probability Letters*, 136:142–145.
- Feldman, B., Martin, E. M., and Skotnes, T. (2012). Big data in healthcare hype and hope. *Dr. Bonnie*, 360:122–125.
- Fielding, K. S. and Head, B. W. (2012). Determinants of young australians’ environmental actions: The role of responsibility attributions, locus of control, knowledge and attitudes. *Environmental Education Research*, 18(2):171–186.
- Fischer, E. and Arnold, S. J. (1990). More than a labor of love: Gender roles and christmas gift shopping. *Journal of consumer research*, 17(3):333–345.
- Foad, H. (2010). Familiarity bias. *Behavioural finance: Investors, corporafions, and markets*, pages 277–294.
- Forbes (2016). Roundup of internet of things forecasts and market estimates, 2016. <https://www.forbes.com/sites/louiscolumbus/2016/11/27/roundup-of-internet-of-things-forecasts-and-market-estimates-2016/#51baf849292d>.
- Forsythe, S. M. and Bailey, A. W. (1996). Shopping enjoyment, perceived time poverty, and time spent shopping. *Clothing and Textiles Research Journal*, 14(3):185–191.
- Fraj, E. and Martinez, E. (2006). Environmental values and lifestyles as determining factors of ecological consumer behaviour: an empirical analysis. *Journal of Consumer Marketing*, 23(3):133–144.
- Freedman, D., Pisani, R., and Purves, R. (1998). *Statistics* (3rd edn).
- Frick, J., Kaiser, F. G., and Wilson, M. (2004). Environmental knowledge and conservation behavior: Exploring prevalence and structure in a representative sample. *Personality and Individual differences*, 37(8):1597–1613.
- Fuchs, V. R. (1980). Time preference and health: an exploratory study.

- Fuchs, V. R. (1991). National health insurance revisited. *Health Affairs*, 10(4):7–17.
- Fuchs, V. R. et al. (1996). Economics, values, and health care reform. *American Economic Review*, 86(1):1–24.
- Fuj, E. T., Hennessy, M., and Mak, J. (1985). An evaluation of the validity and reliability of survey response data on household electricity conservation. *Evaluation Review*, 9(1):93–104.
- Gatersleben, B., Murtagh, N., and Abrahamse, W. (2014). Values, identity and pro-environmental behaviour. *Contemporary Social Science*, 9(4):374–392.
- George, B. (2002). The relationship between lottery ticket and scratch-card buying behaviour, personality and other compulsive behaviours. *Journal of Consumer Behaviour: An International Research Review*, 2(1):7–22.
- Gerbing, D. W., Ahadi, S. A., and Patton, J. H. (1987). Toward a conceptualization of impulsivity: Components across the behavioral and self-report domains. *Multivariate behavioral research*, 22(3):357–379.
- Gervilla, E., Cajal, B., Roca, J., and Palmer, A. (2010). Modelling alcohol consumption during adolescence using zero inflated negative binomial and decision trees. *European Journal of Psychology Applied to Legal Context*, 2(2).
- Gervilla, E. and Palmer, A. (2010). Prediction of cannabis and cocaine use in adolescence using decision trees and logistic regression. *European Journal of Psychology Applied to Legal Context*, 2(1).
- Ghani, R. (2019). In conversations as part of the data science for social good (dssg) fellowship.
- Ghose, A., Ipeirotis, P. G., and Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science*, 31(3):493–520.
- Gifford, R. and Nilsson, A. (2014). Personal and social factors that influence pro-environmental concern and behaviour: A review. *International Journal of Psychology*, 49(3):141–157.
- Gladstone, J. J., Matz, S. C., and Lemaire, A. (2019). Can psychological traits be inferred from spending? evidence from transaction data. *Psychological science*, page 0956797619849435.
- Gladwell, M. (2008). *Outliers: The story of success*. Hachette UK.
- Glaeser, E. L., Kim, H., and Luca, M. (2018). How companies can use the data they collect to further the public good. <https://hbr.org/2018/05/how-companies-can-use-the-data-they-collect-to-further-the-public-good>.

- Gohary, A. and Hanzae, K. H. (2014). Personality traits as predictors of shopping motivations and behaviors: a canonical correlation analysis. *Arab Economic and Business Journal*, 9(2):166–174.
- Goldberg, L. R. (1990). An alternative” description of personality”: the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.
- Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.
- Goldsmith, R. E. and Flynn, L. R. (2015). The etiology of frugal spending: A partial replication and extension. *Comprehensive Psychology*, 4:09–20.
- Goldsmith, R. E., Flynn, L. R., and Clark, R. A. (2014). The etiology of the frugal consumer. *Journal of Retailing and Consumer Services*, 21(2):175–184.
- Gosling, S. D., Rentfrow, P. J., and Potter, J. (2014). Norms for the ten item personality inventory. *Unpublished Data*.
- Gosling, S. D., Rentfrow, P. J., and Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.
- Gould, S. J., Houston, F. S., and Mundt, J. (1997). Failing to try to consume: a reversal of the usual consumer research perspective. *ACR North American Advances*.
- GOV (2015). Plastic bag numbers rise for the fifth year. <https://www.gov.uk/government/news/plastic-bag-numbers-rise-for-the-fifth-year>.
- Greenberg, P. E., Sisitsky, T., Kessler, R. C., Finkelstein, S. N., Berndt, E. R., Davidson, J. R., Ballenger, J. C., and Fyer, A. J. (1999). The economic burden of anxiety disorders in the 1990s. *The Journal of clinical psychiatry*.
- Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4):308–319.
- Guber, D. L. (2003). *The grassroots of a green revolution: Polling America on the environment*. MIT Press.
- Gustavsson, J. P., Jönsson, E. G., Linder, J., and Weinryb, R. M. (2003). The hp5 inventory: definition and assessment of five health-relevant personality traits from a five-factor model perspective. *Personality and Individual Differences*, 35(1):69–89.
- Gutiérrez, J. L. G., Jiménez, B. M., Hernández, E. G., Pcn, C., et al. (2005). Personality and subjective well-being: Big five correlates and demographic variables. *Personality and Individual Differences*, 38(7):1561–1569.
- Hackshaw, A. (2008). Small studies: strengths and limitations.

- Han, Y. K., Morgan, G. A., Kotsiopoulos, A., and Kang-Park, J. (1991). Impulse buying behavior of apparel purchasers. *Clothing and Textiles Research Journal*, 9(3):15–21.
- Hanel, P. H. and Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public? *PloS one*, 11(12):e0168354.
- Haoran, H. (2010). *The effects of an environmental policy on consumers: lessons from the Chinese plastic bag regulation*. Institutionen för nationalekonomi med statistik, Handelshögskolan vid
- Harlow, L. L. and Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological methods*, 21(4):447.
- Harrell Jr, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., and Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2):143–152.
- Harrison, V., Proudfoot, J., Wee, P. P., Parker, G., Pavlovic, D. H., and Manicavasagar, V. (2011). Mobile mental health: review of the emerging field and proof of concept study. *Journal of mental health*, 20(6):509–524.
- Hartig, T., Kaiser, F. G., and Bowler, P. A. (2001). Psychological restoration in nature as a positive motivation for ecological behavior. *Environment and behavior*, 33(4):590–607.
- Hartley, B. L., Thompson, R. C., and Pahl, S. (2015). Marine litter education boosts children’s understanding and self-reported actions. *Marine pollution bulletin*, 90(1-2):209–217.
- Hauser, J. R., Urban, G. L., Liberali, G., and Braun, M. (2009). Website morphing. *Marketing Science*, 28(2):202–223.
- Hausman, A. (2000). A multi-method investigation of consumer motivations in impulse buying behavior. *Journal of consumer marketing*, 17(5):403–426.
- Hayes, N. and Joseph, S. (2003). Big 5 correlates of three measures of subjective well-being. *Personality and Individual differences*, 34(4):723–727.
- He, H. (2010). The effects of an environmental policy on consumers: Lessons from the chinese plastic bag regulation. *rapport nr.: Working Papers in Economics 453*.
- Heine, S. J., Lehman, D. R., Peng, K., and Greenholtz, J. (2002). What’s wrong with cross-cultural comparisons of subjective likert scales?: The reference-group effect. *Journal of personality and social psychology*, 82(6):903.
- Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not weird. *Nature*, 466(7302):29.
- Hermon, R. and Williams, P. A. (2014). Big data in healthcare: What is it used for?

- Hernández, B., Jiménez, J., and José Martín, M. (2011). Age, gender and income: do they really moderate online shopping behaviour? *Online information review*, 35(1):113–133.
- Herrett, E., Gallagher, A. M., Bhaskaran, K., Forbes, H., Mathur, R., van Staa, T., and Smeeth, L. (2015). Data resource profile: clinical practice research datalink (cprd). *International journal of epidemiology*, 44(3):827–836.
- Herring, S. (1996). Posting in a different voice: Gender and ethics in computer-mediated communication. *Philosophical perspectives on computer-mediated communication*, 115:45.
- Hill, G. and Howell, R. T. (2014). Moderators and mediators of pro-social spending and well-being: The influence of values and psychological need satisfaction. *Personality and Individual Differences*, 69:69–74.
- Hines, J. M., Hungerford, H. R., and Tomera, A. N. (1987). Analysis and synthesis of research on responsible environmental behavior: A meta-analysis. *The Journal of environmental education*, 18(2):1–8.
- Hirsh, J. B. (2010). Personality and environmental concern. *Journal of Environmental Psychology*, 30(2):245–248.
- Hirsh, J. B. and Dolderman, D. (2007). Personality predictors of consumerism and environmentalism: A preliminary study. *Personality and individual differences*, 43(6):1583–1593.
- Ho, V. T., Weingart, L. R., and Rousseau, D. M. (2004). Responses to broken promises: does personality matter? *Journal of Vocational Behavior*, 65(2):276–293.
- Hoch, S. J. and Loewenstein, G. F. (1991). Time-inconsistent preferences and consumer self-control. *Journal of consumer research*, 17(4):492–507.
- Hofstede, G. and McCrae, R. R. (2004). Personality and culture revisited: Linking traits and dimensions of culture. *Cross-cultural research*, 38(1):52–88.
- Holbrook, M. B. and Hirschman, E. C. (1982). The experiential aspects of consumption: Consumer fantasies, feelings, and fun. *Journal of consumer research*, 9(2):132–140.
- Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, 92(5):1644–1655.
- Honnold, J. A. (1984). Age and environmental concern some specification of effects. *The Journal of Environmental Education*, 16(1):4–9.
- Howard, G. S., Delgado, E., Miller, D., and Gubbins, S. (1993). Transforming values into actions: Ecological preservation through energy conservation. *The Counseling Psychologist*, 21(4):582–596.
- IBM (2011). Storagenewsletter: Every day we create 2.5 quintillion bytes of data. <http://www.storagenewsletter.com/rubriques/marketreportsresearch/ibmcmstudy/>.

- IBM (2017). 10 key marketing trends for 2017 and ideas for exceeding customer expectations. <https://public.dhe.ibm.com/common/ssi/ecm/wr/en/wr112345usen/watson-customer-engagement-watson-marketing-wr-other-papers-and-reports-wr112345usen-20170719.pdf>.
- Jackson, T. (2012). The challenge of sustainable lifestyles. In *State of the World 2008*, pages 45–60. Routledge.
- Jacobs, B. J., Donkers, B., and Fok, D. (2016). Model-based purchase predictions for large assortments. *Marketing Science*, 35(3):389–404.
- Jakovcevic, A., Steg, L., Mazzeo, N., Caballero, R., Franco, P., Putrino, N., and Favara, J. (2014). Charges for plastic bags: Motivational and behavioral effects. *Journal of Environmental Psychology*, 40:372–380.
- Jambeck, J. R., Geyer, R., Wilcox, C., Siegler, T. R., Perryman, M., Andrady, A., Narayan, R., and Law, K. L. (2015). Plastic waste inputs from land into the ocean. *Science*, 347(6223):768–771.
- John, O. P., Srivastava, S., et al. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.
- Johnson, S. B., Blum, R. W., and Giedd, J. N. (2009). Adolescent maturity and the brain: the promise and pitfalls of neuroscience research in adolescent health policy. *Journal of Adolescent Health*, 45(3):216–221.
- Johnson, T. and Attmann, J. (2009). Compulsive buying in a product specific context: clothing. *Journal of Fashion Marketing and Management: An International Journal*, 13(3):394–405.
- Joireman, J. A., Van Lange, P. A., and Van Vugt, M. (2004). Who cares about the environmental impact of cars? those with an eye toward the future. *Environment and Behavior*, 36(2):187–206.
- Jokela, M. (2009). Personality predicts migration within and between us states. *Journal of Research in Personality*, 43(1):79–83.
- Jokela, M., Bleidorn, W., Lamb, M. E., Gosling, S. D., and Rentfrow, P. J. (2015). Geographically varying associations between personality and life satisfaction in the london metropolitan area. *Proceedings of the National Academy of Sciences*, 112(3):725–730.
- Jones, N. M., Wojcik, S. P., Sweeting, J., and Silver, R. C. (2016). Tweeting negative emotion: An investigation of twitter data in the aftermath of violence on college campuses. *Psychological methods*, 21(4):526.

- Joshanloo, M., Rastegar, P., and Bakhshi, A. (2012). The big five personality domains as predictors of social wellbeing in iranian university students. *Journal of Social and Personal Relationships*, 29(5):639–660.
- Kacen, J. J. and Lee, J. A. (2002). The influence of culture on consumer impulsive buying behavior. *Journal of consumer psychology*, 12(2):163–176.
- Kaiser, F. G., Wölfling, S., and Fuhrer, U. (1999). Environmental attitude and ecological behaviour. *Journal of environmental psychology*, 19(1):1–19.
- Kallus, N. (2014). Predicting crowd behavior with big public data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 625–630. ACM.
- Kalof, L., Dietz, T., Guagnano, G., and Stern, P. C. (2002). Race, gender and environmentalism: The atypical values and beliefs of white men. *Race, Gender & Class*, pages 112–130.
- Kaplan, S. C., Levinson, C. A., Rodebaugh, T. L., Menatti, A., and Weeks, J. W. (2015). Social anxiety and the big five personality traits: The interactive relationship of trust and openness. *Cognitive behaviour therapy*, 44(3):212–222.
- Kavousian, A., Rajagopal, R., and Fischer, M. (2013). Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants’ behavior. *Energy*, 55:184–194.
- Kelloway, E. K. and Francis, L. (2013). Longitudinal research and data analysis.
- Kempf, D. S., Lacznia, R. N., and Smith, R. E. (2006). The effects of gender on processing advertising and product trial information. *Marketing Letters*, 17(1):5–16.
- Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584.
- Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., and Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological methods*, 21(4):507.
- Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217.
- Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., and Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication. *Archives of general psychiatry*, 62(6):593–602.
- Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H.-U., and Kendler, K. S. (1994). Lifetime and 12-month prevalence of dsm-iii-r psychiatric disorders in the united states: results from the national comorbidity survey. *Archives of general psychiatry*, 51(1):8–19.

- Khan, N., Hui, L. H., Chen, T. B., and Hoe, H. Y. (2016). Impulse buying behaviour of generation y in fashion retail. *International Journal of Business and Management*, 11(1):144.
- Kievit, R., Frankenhuys, W. E., Waldorp, L., and Borsboom, D. (2013). Simpson’s paradox in psychological science: a practical guide. *Frontiers in psychology*, 4:513.
- Kim, Y.-M. and Shim, K.-Y. (2002). The influence of internet shopping mall characteristics and user traits on purchase intent. *Irish Marketing Review*, 15:25–34.
- Kline, D. (2014). How many customers does amazon have. *The Motley Fool*.
- Klineberg, S. L., McKeever, M., and Rothenbach, B. (1998). Demographic predictors of environmental concern: It does make a difference how it’s measured. *Social science quarterly*, pages 734–753.
- Kollat, D. T. and Willett, R. P. (1967). Customer impulse purchasing behavior. *Journal of marketing research*, 4(1):21–31.
- Kononenko, I. et al. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(Jan):1–18.
- Korgaonkar, P. K. and Bellenger, D. N. (1980). Profiling the recreational shopper. *Journal of Retailing*, 56(Fall):77–91.
- Kormos, C. and Gifford, R. (2014). The validity of self-report measures of proenvironmental behavior: A meta-analytic review. *Journal of Environmental Psychology*, 40:359–371.
- Kosinski, M. and Behrend, T. (2017). Editorial overview: Big data in the behavioral sciences.
- Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., and Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543.
- Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.
- Kosinski, M., Wang, Y., Lakkaraju, H., and Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological methods*, 21(4):493.
- Koski, N. (2004). Impulse buying on the internet: encouraging and discouraging factors. *Frontiers of E-business Research*, 4:23–35.
- Kotov, R., Gamez, W., Schmidt, F., and Watson, D. (2010). Linking “big” personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychological bulletin*, 136(5):768.

- Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.
- Krause, J., Perer, A., and Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5686–5697. ACM.
- Kroenke, K., Spitzer, R. L., Williams, J. B., Monahan, P. O., and Löwe, B. (2007). Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Annals of internal medicine*, 146(5):317–325.
- Krug, S. E. and Kulhavy, R. W. (1973). Personality differences across regions of the united states. *The Journal of social psychology*, 91(1):73–79.
- Kruger, D. and Byker, D. (2009). Evolved foraging psychology underlies sex differences in shopping experiences and behaviors. *Journal of Social, Evolutionary, and Cultural Psychology*, 3(4):328.
- Kuruvilla, S. J., Joshi, N., and Shah, N. (2009). Do men and women really shop differently? an exploration of gender differences in mall shopping in india. *International Journal of Consumer Studies*, 33(6):715–723.
- Lahey, B. B. (2009). Public health significance of neuroticism. *American Psychologist*, 64(4):241.
- Lakhal, S., Frenette, É., Sévigny, S., and Khechine, H. (2012). Relationship between choice of a business major type (thing-oriented versus person-oriented) and big five personality traits. *The International Journal of Management Education*, 10(2):88–100.
- Lallich, S., Teytaud, O., and Prudhomme, E. (2006). Statistical inference and data mining: false discoveries control. In *Compstat 2006-Proceedings in Computational Statistics*, pages 325–336. Springer.
- Lamiell, J. T. (2015). Statistical thinking in psychological research. *The Wiley handbook of theoretical and philosophical psychology*, pages 200–215.
- Landers, R. N., Brusso, R. C., Cavanaugh, K. J., and Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological methods*, 21(4):475.
- Landis, B. and Gladstone, J. J. (2017). Personality, income, and compensatory consumption: low-income extraverts spend more on status. *Psychological science*, 28(10):1518–1520.
- Lanzini, P. and Thøgersen, J. (2014). Behavioural spillover in the environmental domain: an intervention study. *Journal of Environmental Psychology*, 40:381–390.

- Laroche, M., Saad, G., Cleveland, M., and Browne, E. (2000). Gender differences in information search strategies for a christmas gift. *Journal of consumer marketing*, 17(6):500–522.
- LaRose, R. and Eastin, M. S. (2002). Is online buying out of control? electronic commerce and consumer self-regulation. *Journal of Broadcasting & Electronic Media*, 46(4):549–564.
- Lastovicka, J. L., Bettencourt, L. A., Hughner, R. S., and Kuntze, R. J. (1999). Lifestyle of the tight and frugal: Theory and measurement. *Journal of consumer research*, 26(1):85–98.
- Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205.
- Lebart, L., Morineau, A., and Piron, M. (1995). *Statistique exploratoire multidimensionnelle*, volume 3. Dunod Paris.
- Lee, J. W. (2009). Relationship between consumer personality and brand personality as self-concept: From the case of korean automobile brands. *Academy of Marketing Studies Journal*, 13(1):25–44.
- Leon, A. C., Olfson, M., Broadhead, W. E., Barrett, J. E., Blacklow, R. S., Keller, M. B., Higgins, E. S., and Weissman, M. M. (1995). Prevalence of mental disorders in primary care: implications for screening. *Archives of Family Medicine*, 4(10):857.
- Leonard, J. (2018). Everything you need to know about anxiety medications. <https://www.medicalnewstoday.com/articles/323666.php>.
- Leskovec, J. and Krevl, A. (2014). Snap: Stanford network analysis project.
- Levy, M. F. (1976). Deferred gratification and social class. *The Journal of Social Psychology*, 100(1):123–135.
- Lewinsohn, P. M., Gotlib, I. H., Lewinsohn, M., Seeley, J. R., and Allen, N. B. (1998). Gender differences in anxiety disorders and anxiety symptoms in adolescents. *Journal of abnormal psychology*, 107(1):109.
- Li, J. and Wang, Z. J. (2009). Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. *Journal of Machine Learning Research*, 10(Feb):475–514.
- Lin, M., Lucas Jr, H. C., and Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4):906–917.
- Linden, G., Smith, B., and York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, (1):76–80.
- Loftus, E. F. and Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of verbal learning and verbal behavior*, 13(5):585–589.

- Longo, Y., Coyne, I., Joseph, S., and Gustavsson, P. (2016). Support for a general factor of well-being. *Personality and Individual Differences*, 100:68–72.
- Louis, D. and Lombart, C. (2010). Impact of brand personality on three major relational consequences (trust, attachment, and commitment to the brand). *Journal of Product & Brand Management*, 19(2):114–130.
- Lu, L. (1997). Social support, reciprocity, and well-being. *The Journal of social psychology*, 137(5):618–628.
- Lu, S., Xiao, L., and Ding, M. (2016). A video-based automated recommender (var) system for garments. *Marketing Science*, 35(3):484–510.
- Lubinski, D. and Humphreys, L. G. (1990). Assessing spurious” moderator effects”: Illustrated substantively with the hypothesized (“ synergistic”) relation between spatial and mathematical ability. *Psychological bulletin*, 107(3):385.
- Ludeke, S. G. and Larsen, E. G. (2017). Problems with the big five assessment in the world values survey. *Personality and Individual Differences*, 112:103–105.
- Luminet, O., Zech, E., Rimé, B., and Wagner, H. (2000). Predicting cognitive and social consequences of emotional episodes: The contribution of emotional intensity, the five factor model, and alexithymia. *Journal of Research in Personality*, 34(4):471–497.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.
- Luo, X. (2005). How does shopping with others influence impulsive purchasing? *Journal of Consumer psychology*, 15(4):288–294.
- Lynn, R. and Martin, T. (1997). Gender differences in extraversion, neuroticism, and psychoticism in 37 nations. *The Journal of social psychology*, 137(3):369–373.
- Mahmoodi, J., Leckelt, M., van Zalk, M. W., Geukes, K., and Back, M. D. (2017). Big data approaches in social and behavioral science: four key trade-offs and a call for integration. *Current Opinion in Behavioral Sciences*, 18:57–62.
- Mann, J. J., Ellis, S. P., Waternaux, C. M., Liu, X., Oquendo, M. A., Malone, K. M., Brodsky, B. S., Haas, G. L., and Currier, D. (2008). Classification trees distinguish suicide attempters in major psychiatric disorders: a model of clinical decision making. *The Journal of clinical psychiatry*, 69(1):23.
- Marcus, B., Machilek, F., and Schütz, A. (2006). Personality in cyberspace: personal web sites as media for personality expressions and impressions. *Journal of personality and social psychology*, 90(6):1014.

- Markowitz, E. M., Goldberg, L. R., Ashton, M. C., and Lee, K. (2012). Profiling the “pro-environmental individual”: A personality perspective. *Journal of personality*, 80(1):81–111.
- Mathews, A. and Ridgeway, V. (1981). Personality and surgical recovery: a review. *British Journal of Clinical Psychology*, 20(4):243–260.
- Matz, S. C., Gladstone, J. J., and Stillwell, D. (2016). Money buys happiness when spending fits our personality. *Psychological science*, 27(5):715–725.
- Matz, S. C., Gladstone, J. J., and Stillwell, D. (2017). In a world of big data, small effects can still matter: a reply to boyce, daly, hounkpatin, and wood (2017). *Psychological science*, 28(4):547–550.
- Matz, S. C. and Netzer, O. (2017). Using big data as a window into consumers’ psychology. *Current opinion in behavioral sciences*, 18:7–12.
- Matzler, K., Bidmon, S., and Grabner-Kräuter, S. (2006). Individual determinants of brand affect: the role of the personality traits of extraversion and openness to experience. *Journal of Product & Brand Management*, 15(7):427–434.
- Matzler, K., Grabner-Kräuter, S., and Bidmon, S. (2008). Risk aversion and brand loyalty: the mediating role of brand trust and brand affect. *Journal of Product & Brand Management*, 17(3):154–162.
- McCrae, R. R. (1993). Openness to experience as a basic dimension of personality. *Imagination, Cognition and Personality*, 13(1):39–55.
- McCrae, R. R. and Costa, P. T. (2003). *Personality in adulthood: A five-factor theory perspective*. Guilford Press.
- McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.
- McCullagh, P. and Nelder, J. (1989). Generalized linear models., 2nd edn.(chapman and hall: London). *Standard book on generalized linear models*.
- McFadden, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.
- McFatter, R. M. (1994). Interactions in predicting mood from extraversion and neuroticism. *Journal of personality and social psychology*, 66(3):570.
- McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.
- Meyers-Levy, J. and Sternthal, B. (1991). Gender differences in the use of message cues and judgments. *Journal of marketing research*, 28(1):84–96.

- Miah, S. J., Vu, H. Q., Gammack, J., and McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. *Information & Management*, 54(6):771–785.
- Mick, D. G. and DeMoss, M. (1990). To me from me: A descriptive phenomenology of self-gifts. *ACR North American Advances*.
- Milfont, T. L. and Sibley, C. G. (2012). The big five personality traits and environmental engagement: Associations at the individual and societal level. *Journal of Environmental Psychology*, 32(2):187–195.
- Minitab (2013). What are the effects of multicollinearity and when can i ignore them? <https://blog.minitab.com/blog/adventures-in-statistics-2/what-are-the-effects-of-multicollinearity-and-when-can-i-ignore-them>.
- Mischel, W., Shoda, Y., and Peake, P. K. (1988). The nature of adolescent competencies predicted by preschool delay of gratification. *Journal of personality and social psychology*, 54(4):687.
- Molloy, G., O’carroll, R., and Ferguson, E. (2013). Conscientiousness and medication adherence: a meta-analysis. *Annals of Behavioral Medicine*, 47(1):92–101.
- Monaghan, A. (2018). Which items can’t be recycled? <https://www.theguardian.com/environment/2018/mar/16/which-items-cant-be-recycled-plastic-bags-coffee-cups>.
- Mooradian, T. A. and Olver, J. M. (1997). “i can’t get no satisfaction:” the impact of personality and emotion on postpurchase processes. *Psychology & Marketing*, 14(4):379–393.
- Moras, K., Clark, L., Katon, W., Roy-Byrne, R., Watson, D., and Barlow, D. (1996). Mixed anxiety-depression. *DSM-IV sourcebook*, 2:623–643.
- Morineau, A., Rakotomalala, R., and MODULAD, P. (2006). Critère vt100 de sélection des règles d’association. In *EGC*, pages 581–592.
- Mowen, J. C. and Spears, N. (1999). Understanding compulsive buying among college students: A hierarchical approach. *Journal of Consumer Psychology*, 8(4):407–430.
- Muris, P., Bos, A. E., Mayer, B., Verkade, R., Thewissen, V., and Dell’Avvento, V. (2009). Relations among behavioral inhibition, big five personality factors, and anxiety disorder symptoms in non-clinical children. *Personality and Individual Differences*, 46(4):525–529.
- Murray, D. R. and Schaller, M. (2014). Pathogen prevalence and geographical variation in traits and behavior. *Psychological geography*, pages 51–70.
- Nagra, G. and Gopal, R. (2013). An study of factors affecting on online shopping behavior of consumers. *International journal of scientific and research publications*, 3(6):1–4.

- Nechay, D., Pointurier, Y., and Coates, M. (2009). Controlling false alarm/discovery rates in online internet traffic flow classification. In *IEEE INFOCOM 2009*, pages 684–692. IEEE.
- NEFE (2010). Spendster contest offers redemption for spenders willing to confess. <http://www.nefe.org/press-room/news/spendster-contest-offers-redemption>.
- NEMA (2018). Ban on plastic carrier bags. https://www.nema.go.ke/index.php?option=com_content&view=article&id=102:government-bans-plastic-carriers&catid=10&Itemid=246.
- Nesbitt, S. (1959). Today’s housewives plan menus as they shop. *Nesbitt associates release*, 2(3).
- Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3):521–543.
- New, J., Krasnow, M. M., Truxaw, D., and Gaulin, S. J. (2007). Spatial adaptations for plant foraging: Women excel and calories count. *Proceedings of the Royal Society B: Biological Sciences*, 274(1626):2679–2684.
- Ng, A. Y. (2004). Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM.
- Nguyen, T. T., Harper, F. M., Terveen, L., and Konstan, J. A. (2018). User personality and user satisfaction with recommender systems. *Information Systems Frontiers*, 20(6):1173–1189.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.
- Nielsen, T. D. and Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.
- Noë, N., Whitaker, R. M., and Allen, S. M. (2018). Personality homophily and geographic distance in facebook. *Cyberpsychology, Behavior, and Social Networking*, 21(6):361–366.
- Nosek, B. A., Spies, J. R., and Motyl, M. (2012). Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631.
- NVP (2018). How big data and ai are driving business innovation. <http://newvantage.com/wp-content/uploads/2018/01/Big-Data-Executive-Survey-2018-Findings-1.pdf>.
- O’Dea, J. A. and Caputi, P. (2001). Association between socioeconomic status, weight, age and gender, and the body image and weight control practices of 6-to 19-year-old children and adolescents. *Health education research*, 16(5):521–532.

- Oh, J., Laubach, M., and Luczak, A. (2003). Estimating neuronal variable importance with random forest. In *2003 IEEE 29th Annual Proceedings of Bioengineering Conference*, pages 33–34. IEEE.
- Olfson, M., Shea, S., Feder, A., Fuentes, M., Nomura, Y., Gameroff, M., and Weissman, M. M. (2000). Prevalence of anxiety, depression, and substance use disorders in an urban general medicine practice. *Archives of family medicine*, 9(9):876.
- Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis*, 27(4):286–306.
- Ormel, J., VonKorff, M., Ustun, T. B., Pini, S., Korten, A., and Oldehinkel, T. (1994). Common mental disorders and disability across cultures: results from the who collaborative study on psychological problems in general health care. *Jama*, 272(22):1741–1748.
- Oswald, F. L. and Putka, D. J. (2015). Statistical methods for big data: A scenic tour. In *Big Data at Work*, pages 57–77. Routledge.
- O’Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5):673–690.
- O’Connor, F. (2014). Health-it early adopters well poised for big-data advances in clinical medicine.[internet] 2013.[cited march 17, 2014].
- Parboteeah, D. V. (2005). A model of online impulse buying: An empirical study.
- Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., and Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.
- Paul, J., Verleysen, M., Dupont, P., et al. (2013). Identification of statistically significant features from random forests. In *ECML workshop on Solving Complex Machine Learning Problems with Ensemble Methods*, pages 69–80.
- Paunonen, S. V. (1998). Hierarchical organization of personality and prediction of behavior. *Journal of personality and social psychology*, 74(2):538.
- Paunonen, S. V. (2003). Big five factors of personality and replicated predictions of behavior. *Journal of personality and social psychology*, 84(2):411.
- Pawluk, E. J. and Koerner, N. (2013). A preliminary investigation of impulsivity in generalized anxiety disorder. *Personality and Individual Differences*, 54(6):732–737.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

- Peña, J. M. (2008). Learning gaussian graphical models of gene networks with false discovery rate control. In *European conference on evolutionary computation, machine learning and data mining in bioinformatics*, pages 165–176. Springer.
- Pennebaker, J. W. and King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.
- Pepper, M., Jackson, T., and Uzzell, D. (2009). An examination of the values that motivate socially conscious and frugal consumer behaviours. *International journal of consumer studies*, 33(2):126–136.
- Pepper, M., Jackson, T., and Uzzell, D. (2011). An examination of christianity and socially conscious and frugal consumer behaviors. *Environment and Behavior*, 43(2):274–290.
- Petrov, C. (2014). With big data comes big responsibility. <https://hbr.org/2014/11/with-big-data-comes-big-responsibility>.
- Petrov, C. (2019). Big data statistics 2019. <https://techjury.net/stats-about/big-data-statistics/>.
- Petrovici, D. A. and Ritson, C. (2006). Factors influencing consumer dietary health preventative behaviours. *BMC Public Health*, 6(1):222.
- Pettus, A. M. and Giles, M. B. (1987). Personality characteristics and environmental attitudes. *Population and Environment*, 9(3):127–137.
- Piron, F. (1991). Defining impulse purchasing. *ACR North American Advances*.
- Plaut, V. C., Markus, H. R., and Lachman, M. E. (2002). Place matters: Consensual features and regional variation in american well-being and self. *Journal of personality and social psychology*, 83(1):160.
- Podoshen, J. S., Andrzejewski, S. A., and Hunt, J. M. (2014). Materialism, conspicuous consumption, and american hip-hop subculture. *Journal of International Consumer Marketing*, 26(4):271–283.
- Poortinga, W., Sautkina, E., Thomas, G. O., and Wolstenholme, E. (2016). The english plastic bag charge: Changes in attitudes and behaviour.
- Poortinga, W., Whitmarsh, L., and Suffolk, C. (2013). The introduction of a single-use carrier bag charge in wales: Attitude change and behavioural spillover effects. *Journal of Environmental Psychology*, 36:240–247.
- Popper, K. R. (1949). On the theory of deduction, part i. derivation and its generalizations.
- Proto, E. and Rustichini, A. (2015). Life satisfaction, income and personality. *Journal of Economic Psychology*, 48:17–32.

- Puri, R. (1996). Measuring and modifying consumer impulsiveness: A cost-benefit accessibility framework. *Journal of consumer Psychology*, 5(2):87–113.
- Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 180–185. IEEE.
- Quevedo, R. J. M. and Abella, M. C. (2011). Well-being and personality: Facet-level analyses. *Personality and Individual Differences*, 50(2):206–211.
- Rammstedt, B. and John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212.
- Regier, D. A., Narrow, W. E., and Rae, D. S. (1990). The epidemiology of anxiety disorders: The epidemiologic catchment area (eca) experience. *Journal of Psychiatric Research*, 24:3–14.
- Reimers, S., Maylor, E. A., Stewart, N., and Chater, N. (2009). Associations between a one-shot delay discounting measure and age, income, education and real-world impulsive behavior. *Personality and Individual Differences*, 47(8):973–978.
- Rentfrow, P. J., Gosling, S. D., Jokela, M., Stillwell, D. J., Kosinski, M., and Potter, J. (2013). Divided we stand: Three psychological regions of the united states and their political, economic, social, and health correlates. *Journal of personality and social psychology*, 105(6):996.
- Rentfrow, P. J., Gosling, S. D., and Potter, J. (2008). A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics. *Perspectives on Psychological Science*, 3(5):339–369.
- Rentfrow, P. J., Jokela, M., and Lamb, M. E. (2015). Regional personality differences in great britain. *PLoS One*, 10(3):e0122245.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.
- Richins, M. L. and Dawson, S. (1992). A consumer values orientation for materialism and its measurement: Scale development and validation. *Journal of consumer research*, 19(3):303–316.
- Rick, S. I., Cryder, C. E., and Loewenstein, G. (2007). Tightwads and spendthrifts. *Journal of consumer research*, 34(6):767–782.

- Ringel, D. M. and Skiera, B. (2014). Understanding competition using big consumer search data. In *2014 47th Hawaii International Conference on System Sciences*, pages 3129–3138. IEEE.
- Roberts, B. W., Walton, K. E., and Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological bulletin*, 132(1):1.
- Romal, J. B. and Kaplan, B. J. (1995). Difference in self-control among spenders and savers. *Psychology: A journal of human behavior*.
- Rook, D. W. and Fisher, R. J. (1995). Normative influences on impulsive buying behavior. *Journal of consumer research*, 22(3):305–313.
- Rook, D. W. and Hoch, S. J. (1985). Consuming impulses. *ACR North American Advances*.
- Roosendaal, B., McEwen, B. S., and Chattarji, S. (2009). Stress, memory and the amygdala. *Nature Reviews Neuroscience*, 10(6):423.
- Rose, P., Smith, S. T., and Segrist, D. J. (2010). Too cheap to chug: frugality as a buffer against college-student drinking. *Journal of Consumer Behaviour*, 9(3):228–238.
- Rossi, A., Amaddeo, F., Sandri, M., and Tansella, M. (2005). Determinants of once-only contact in a community-based psychiatric service. *Social Psychiatry and Psychiatric Epidemiology*, 40(1):50–56.
- Rothchild, I. (2006). Induction, deduction, and the scientific method. *Soc. study Reprod.*
- Royer, S.-J., Ferron, S., Wilson, S. T., and Karl, D. M. (2018). Production of methane and ethylene from plastic in the environment. *PloS one*, 13(8):e0200574.
- Rucker, D. D. and Galinsky, A. D. (2009). Conspicuous consumption versus utilitarian ideals: How different levels of power shape consumer behavior. *Journal of Experimental Social Psychology*, 45(3):549–555.
- Ruhm, C. J. (2012). Understanding overeating and obesity. *Journal of Health economics*, 31(6):781–796.
- Rushton, J. P. and Bons, T. A. (2005). Mate choice and friendship in twins: evidence for genetic similarity. *Psychological Science*, 16(7):555–559.
- Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., and Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.
- Saucier, G. and Goldberg, L. R. (1998). What is beyond the big five? *Journal of personality*, 66:495–524.
- Saucier, G. and Goldberg, L. R. (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of personality*, 69(6):847–879.

- Schaller, M. (2006). Parasites, behavioral defenses, and the social psychological mechanisms through which cultures are evoked. *Psychological Inquiry*, 17(2):96–101.
- Schmutte, P. S. and Ryff, C. D. (1997). Personality and well-being: reexamining methods and meanings. *Journal of Personality and Social Psychology*, 73(3):549.
- Schultze, U. and Mason, R. O. (2012). Studying cyborgs: re-examining internet studies as human subjects research. *Journal of Information Technology*, 27(4):301–312.
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.
- Scottish Government (2015). Scotland’s single use carrier bag usage drops by 650 million since charge was introduced. <https://www.webarchive.org.uk/wayback/archive/20181023201037/https://news.gov.scot/news/bags-of-success>.
- Secchi, P. (2018). On the role of statistics in the era of big data: A call for a debate. *Statistics & Probability Letters*, 136:10–14.
- Sharp, A., Høj, S., and Wheeler, M. (2010). Proscription and its impact on anti-consumption behaviour and attitudes: the case of plastic bags. *Journal of Consumer Behaviour*, 9(6):470–484.
- Sharwood (2017). Developing world hits 98.7 per cent mobile phone adoption. https://www.theregister.co.uk/2017/08/03/itu_facts_and_figures_2017/.
- Shen, K.-Q., Ong, C.-J., Li, X.-P., Hui, Z., and Wilder-Smith, E. P. (2007). A feature selection method for multilevel mental fatigue eeg classification. *IEEE transactions on biomedical engineering*, 54(7):1231–1237.
- Sheth, J. N. and Parvatiyar, A. (1995). The evolution of relationship marketing. *International business review*, 4(4):397–418.
- Shmueli, G. et al. (2010). To explain or to predict? *Statistical science*, 25(3):289–310.
- Shoham, A. and Brenčič, M. M. (2004). Value, price consciousness, and consumption frugality: An empirical study. *Journal of International Consumer Marketing*, 17(1):55–69.
- Shrout, P. E. and Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology*, 69:487–510.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

- Sirgy, M. J. (1982). Self-concept in consumer behavior: A critical review. *Journal of consumer research*, 9(3):287–300.
- Sirgy, M. J. (1985). Using self-congruity and ideal congruity to predict purchase motivation. *Journal of business Research*, 13(3):195–206.
- Sivanathan, N. and Pettit, N. C. (2010). Protecting the self through consumption: Status goods as affirmational commodities. *Journal of Experimental Social Psychology*, 46(3):564–570.
- Skatova, A. and Goulding, J. (2019). Psychology of personal data donation.
- Skatova, A., Johal, J., Houghton, R., Mortier, R., Bhandari, N., Lodge, T., Wagner, C., Goulding, J., Crowcroft, J., and Madhavapeddy, A. (2013). Perceived risks of personal data sharing. *Proc. Digital Economy: Open Digital*.
- Skatova, A., Ng, E., and Goulding, J. (2014). Data donation: Sharing personal data for public good. *Application of Digital Innovation. London, England: N-Lab*.
- Smedslund, J. (2016). Why psychology cannot be an empirical science. *Integrative Psychological and Behavioral Science*, 50(2):185–195.
- Smith, S. M. (1994). Theoretical principles of context-dependent memory. *Theoretical aspects of memory*, 2:168–195.
- Smithers, R. (2016). England’s plastic bag usage drops 85% since 5p charge introduced. <https://www.theguardian.com/environment/2016/jul/30/england-plastic-bag-usage-drops-85-per-cent-since-5p-charged-introduced>.
- Solnick, J. V., Kannenberg, C. H., Eckerman, D. A., and Waller, M. B. (1980). An experimental analysis of impulsivity and impulse control in humans. *Learning and motivation*, 11(1):61–77.
- Solomon, M., Russell-Bennett, R., and Previte, J. (2012). *Consumer behaviour*. Pearson Higher Education AU.
- Sønderlund, A., Smith, J., Hutton, C., and Kapelan, Z. (2014). Using smart meters for household water consumption feedback: knowns and unknowns.
- Specht, J., Egloff, B., and Schmukle, S. C. (2011). Stability and change of personality across the life course: The impact of age and major life events on mean-level and rank-order stability of the big five. *Journal of personality and social psychology*, 101(4):862.
- Spence, A., Poortinga, W., and Pidgeon, N. (2012). The psychological distance of climate change. *Risk Analysis: An International Journal*, 32(6):957–972.

- Spitzer, R. L., Williams, J. B., Kroenke, K., Linzer, M., Verloin deGruy, F., Hahn, S. R., Brody, D., and Johnson, J. G. (1994). Utility of a new procedure for diagnosing mental disorders in primary care: the prime-md 1000 study. *Jama*, 272(22):1749–1756.
- Stallman, H. M. (2010). Psychological distress in university students: A comparison with general population data. *Australian Psychologist*, 45(4):249–257.
- Stanley, C. and Byrne, M. D. (2016). Comparing vector-based and bayesian memory models using large-scale datasets: User-generated hashtag and tag prediction on twitter and stack overflow. *Psychological methods*, 21(4):542.
- Statistica (2018). Global big data market year-on-year growth from 2012 to 2027. <https://www.statista.com/statistics/255968/global-big-data-market-growth-forecast/>.
- Stevenson, B. and Wolfers, J. (2013). Subjective well-being and income: Is there any evidence of satiation? *American Economic Review*, 103(3):598–604.
- Strahilevitz, M. and Myers, J. G. (1998). Donations to charity as purchase incentives: How well they work may depend on what you are trying to sell. *Journal of consumer research*, 24(4):434–446.
- Strobel, M. (2008). Estimation of minimum mean squared error with variable metric from censored observations.
- Strobl, C. and Zeileis, A. (2008). Danger: High power!—exploring the statistical properties of a test for random forest variable importance.
- Sun, Y., Wang, S., Li, J., Zhao, D., and Fan, J. (2017). Understanding consumers’ intention to use plastic bags: using an extended theory of planned behaviour model. *Natural Hazards*, 89(3):1327–1342.
- Sung, Y. and Kim, J. (2010). Effects of brand personality on brand trust and brand affect. *Psychology & Marketing*, 27(7):639–661.
- Sutter, M., Kocher, M. G., Glätzle-Rützler, D., and Trautmann, S. T. (2013). Impatience and uncertainty: Experimental decisions predict adolescents’ field behavior. *American Economic Review*, 103(1):510–31.
- Swami, V., Chamorro-Premuzic, T., Snelgar, R., and Furnham, A. (2011). Personality, individual differences, and demographic antecedents of self-reported household waste management behaviours. *Journal of Environmental Psychology*, 31(1):21–26.
- Taft, R. (1954). Selective recall and memory distortion of favorable and unfavorable material. *The Journal of Abnormal and Social Psychology*, 49(1):23.
- Tangney, J. P., BOONE, A. L., and BAUMEISTER, R. F. (2004a). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. In *Self-Regulation and Self-Control*, pages 181–220. Routledge.

- Tangney, J. P., BOONE, A. L., and BAUMEISTER, R. F. (2004b). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. In *Self-Regulation and Self-Control*, pages 181–220. Routledge.
- Tapia-Fonllem, C., Corral-Verdugo, V., Fraijo-Sing, B., and Durón-Ramos, M. (2013). Assessing sustainable behavior and its correlates: A measure of pro-ecological, frugal, altruistic and equitable actions. *Sustainability*, 5(2):711–723.
- Tarpey, T. (2000). A note on the prediction sum of squares statistic for restricted least squares. *The American Statistician*, 54(2):116–118.
- Terracciano, A. and Costa Jr, P. T. (2004). Smoking and the five-factor model of personality. *Addiction*, 99(4):472–481.
- American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders (5th ed.). *DSM-IV sourcebook*.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.
- University of Cambridge (2015). Bbc big personality test, 2009-2011.
- Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- Thøgersen, J. and Ölander, F. (2003). Spillover of environment-friendly consumer behaviour. *Journal of environmental psychology*, 23(3):225–236.
- Thomas, G. O., Poortinga, W., and Sautkina, E. (2016). The welsh single-use carrier bag charge and behavioural spillover. *Journal of Environmental Psychology*, 47:126–135.
- Thomas, G. O., Sautkina, E., Poortinga, W., Wolstenholme, E., and Whitmarsh, L. (2019). The english plastic bag charge changed behavior and increased support for other charges to reduce plastic waste. *Frontiers in Psychology*, 10.
- Thompson, C. J., Locander, W. B., and Pollio, H. R. (1990). The lived meaning of free choice: An existential-phenomenological description of everyday consumer experiences of contemporary married women. *Journal of consumer research*, 17(3):346–361.
- Thompson, D. (2009). Ranking predictors in logistic regression. *Paper D10-2009. On-line available at <http://www.mwsug.org/proceedings/2009/stats/MWSUG-2009-D10.pdf>*. (visited 2015, June 25).
- Tifferet, S. and Herstein, R. (2012). Gender differences in brand commitment, impulse buying, and hedonic consumption. *Journal of Product & Brand Management*, 21(3):176–182.

- Tinne, W. S. (2010). Impulse purchasing: A literature overview. *ASA University Review*, 4(2):65–73.
- Tkalcic, M., Kunaver, M., Tasic, J., and Košir, A. (2009). Personality based user similarity measure for a collaborative recommender system. In *Proceedings of the 5th Workshop on Emotion in Human-Computer Interaction-Real world challenges*, pages 30–37.
- Todd, S. and Lawson, R. (2003). Towards an understanding of frugal consumers. *Australasian Marketing Journal (AMJ)*, 11(3):8–18.
- Tsao, W.-C. and Chang, H.-R. (2010). Exploring the impact of personality traits on online shopping behavior. *African Journal of Business Management*, 4(9):1800–1812.
- Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.
- Tversky, A. and Kahneman, D. (1986). Judgment under uncertainty: Heuristics and biases. *Judgment and decision making: An interdisciplinary reader*, pages 38–55.
- Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.
- Uher, J. (2018). Quantitative data from rating scales: An epistemological and methodological enquiry. *Frontiers in psychology*, 9:2599.
- Uher, J., Werner, C. S., and Gosselt, K. (2013). From observations of individual behaviour to social representations of personality: Developmental pathways, attribution biases, and limitations of questionnaire methods. *Journal of Research in Personality*, 47(5):647–667.
- Urien, B. and Kilbourne, W. (2011). Generativity and self-enhancement values in eco-friendly behavioral intentions and environmentally responsible consumption behavior. *Psychology & marketing*, 28(1):69–90.
- Valence, G., d’Astous, A., and Fortier, L. (1988). Compulsive buying: Concept and measurement. *Journal of consumer policy*, 11(4):419–433.
- Van Boven, L. and Gilovich, T. (2003). To do or to have? that is the question. *Journal of personality and social psychology*, 85(6):1193.
- Van de Mortel, T. F. et al. (2008). Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, The*, 25(4):40.
- Van de Vliert, E. (2013). Climato-economic habitats support patterns of human needs, stresses, and freedoms. *Behavioral and Brain Sciences*, 36(5):465–480.
- Van der Linden, S., Maibach, E., and Leiserowitz, A. (2015). Improving public engagement with climate change: Five “best practice” insights from psychological science. *Perspectives on Psychological Science*, 10(6):758–763.

- Van der Pol, M. (2011). Health, education and time preference. *Health economics*, 20(8):917–929.
- Veenhoven (1984). The concept of happiness. In *Conditions of Happiness*, pages 12–38.
- Verplanken, B. and Herabadi, A. (2001). Individual differences in impulse buying tendency: Feeling and no thinking. *European Journal of personality*, 15(S1):S71–S83.
- Verplanken, B., Herabadi, A. G., Perry, J. A., and Silvera, D. H. (2005). Consumer style and health: The role of impulsive buying in unhealthy eating. *Psychology & Health*, 20(4):429–441.
- Virvilaite, R., Tumasonyte, D., and Sliburyte, L. (2015). The influence of word of mouth communication on brand equity: Receiver perspectives. *Procedia-Social and Behavioral Sciences*, 213:641–646.
- Vittersø, J. (2001). Personality traits and subjective well-being: Emotional stability, not extraversion, is probably the important predictor. *Personality and Individual Differences*, 31(6):903–914.
- Vizard, S. (2018). Despite gdpr, consumers still don’t understand how brands use their data. <https://www.boots-uk.com/about-boots-uk/about-boots/boots-in-numbers/>.
- Wagner, C. H. (1982). Simpson’s paradox in real life. *The American Statistician*, 36(1):46–48.
- Walker, R. E., Fryer, C. S., Butler, J., Keane, C. R., Kriska, A., and Burke, J. G. (2011). Factors influencing food buying practices in residents of a low-income food desert and a low-income food oasis. *Journal of Mixed Methods Research*, 5(3):247–267.
- Wang, M. and Cheng, Y. (2013). The sweet burden: Young consumers’ extroversion, number of friendship on online social network and luxury consumption tendency. In *2013 10th International Conference on Service Systems and Service Management*, pages 753–756. IEEE.
- Ward, E. A. (1993). Generalizability of psychological research from undergraduates to employed adults. *The Journal of Social Psychology*, 133(4):513–519.
- Ward, M. R. and Lee, M. J. (2000). Internet shopping, consumer search and product branding. *Journal of product & brand management*, 9(1):6–20.
- Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, 38(4):319–350.
- Wedel, M. and Kannan, P. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, 80(6):97–121.

- Welsh Government (2016). Northern ireland carrier bag levy statistics. https://gov.wales/topics/environmentcountryside/epq/waste_recycling/substance/carrierbags/?lang=en.
- Wenzel, M., von Versen, C., Hirschmüller, S., and Kubiak, T. (2015). Curb your neuroticism—mindfulness mediates the link between neuroticism and subjective well-being. *Personality and Individual Differences*, 80:68–75.
- Westenhoefer, J. (2005). Age and gender dependent profile of food choice. In *Diet diversification and health promotion*, volume 57, pages 44–51. Karger Publishers.
- Westerterp, K. and Goris, A. (2002). Validity of the assessment of dietary intake: Problems of misreporting. *Current opinion in clinical nutrition and metabolic care*, 5:489–93.
- Weston, S. J., Hill, P. L., and Jackson, J. J. (2015). Personality traits predict the onset of disease. *Social Psychological and Personality Science*, 6(3):309–317.
- Weun, S., Jones, M. A., and Beatty, S. E. (1998). Development and validation of the impulse buying tendency scale. *Psychological reports*, 82(3_suppl):1123–1133.
- Whiteside, S. P. and Lynam, D. R. (2003). Understanding the role of impulsivity and externalizing psychopathology in alcohol abuse: application of the upps impulsive behavior scale. *Experimental and clinical psychopharmacology*, 11(3):210.
- Williams, L., O’Connor, R. C., Grubb, N., and O’Carroll, R. (2011). Type d personality predicts poor medication adherence in myocardial infarction patients. *Psychology & health*, 26(6):703–712.
- Williams, S. (2018). Making ai human again: The importance of explainable ai (xai). quantum black. [://medium.com/@QuantumBlack/making-ai-human-again-the-importance-of-explainable-ai-xai-95d347ccbb1c](https://medium.com/@QuantumBlack/making-ai-human-again-the-importance-of-explainable-ai-xai-95d347ccbb1c).
- Wiseman, M. and Bogner, F. X. (2003). A higher-order model of ecological values and its relationship to personality. *Personality and Individual differences*, 34(5):783–794.
- Witkowski, T. H. (2003). World war ii poster campaigns—preaching frugality to american consumers. *Journal of advertising*, 32(1):69–82.
- Wolfe, R. N. and Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational and psychological measurement*, 55(2):177–185.
- Wolfstein, M. and Trull, T. J. (1997). Depression and openness to experience. *Journal of Personality Assessment*, 69(3):614–632.
- Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82.

- Wood, M. (1998). Socio-economic status, delay of gratification, and impulse buying. *Journal of economic psychology*, 19(3):295–320.
- WRAP (2015). Wrap announces new carrier bag figures. http://www.wrap.org.uk/2015_carrier_bag_figures.
- Xanthos, D. and Walker, T. R. (2017). International policies to reduce plastic marine pollution from single-use plastics (plastic bags and microbeads): a review. *Marine pollution bulletin*, 118(1-2):17–26.
- Yarkoni, T. and Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.
- Yawn, B. P., Barrette, B. A., and Wollan, P. C. (2004). Ovarian cancer: the neglected diagnosis. In *Mayo Clinic Proceedings*, volume 79, pages 1277–1282. Elsevier.
- Yik, M. S., Russell, J. A., Ahn, C.-K., Dols, J. M. F., and Suzuki, N. (2002). Relating the five-factor model of personality to a circumplex model of affect. In *The five-factor model of personality across cultures*, pages 79–104. Springer.
- Yoon, S.-H., Park, J.-S., Kim, M.-S., Lim, C., and Cho, J. (2014). Behavior signature for big data traffic identification. In *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, pages 261–266. IEEE.
- Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.
- Youyou, W., Stillwell, D., Schwartz, H. A., and Kosinski, M. (2017). Birds of a feather do flock together: Behavior-based personality-assessment method reveals personality similarity among couples and friends. *Psychological science*, 28(3):276–284.
- Zajonc, R. B. (2001). Mere exposure: A gateway to the subliminal. *Current directions in psychological science*, 10(6):224–228.
- Zelezny, L. C., Chua, P.-P., and Aldrich, C. (2000). New ways of thinking about environmentalism: Elaborating on gender differences in environmentalism. *Journal of Social issues*, 56(3):443–457.
- Zellner, D. A., Saito, S., and Gonzalez, J. (2007). The effect of stress on men’s food selection. *Appetite*, 49(3):696–699.
- Zeng, Z., Pantic, M., and Huang, T. S. (2009). Emotion recognition based on multimodal information. In *Affective Information Processing*, pages 241–265. Springer.
- Zhang, J. W., Howell, R. T., Caprariello, P. A., and Guevarra, D. A. (2014). Damned if they do, damned if they don’t: Material buyers are not happier from material or experiential consumption. *Journal of Research in Personality*, 50:71–83.

- Zhou, K. and Yang, S. (2016). Understanding household energy consumption behavior: The contribution of energy big data analytics. *Renewable and Sustainable Energy Reviews*, 56:810–819.
- Zurawicki, L. (2010). Neural bases for segmentation and positioning. In *Neuromarketing*, pages 163–210. Springer.
- Zwaan, R. A., Etz, A., Lucas, R. E., and Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41.

Chapter 10

Supplementary Materials

10.1 Chapter 3: Methodology and Data

10.1.1 The Personality Questionnaire in Full

The full questionnaire used in this study can be viewed and downloaded via this link: https://neodem.wp.horizon.ac.uk/wp-content/uploads/2018/11/psychology_survey.pdf

10.2 Chapter 4: Personality and Consumer Behaviour Predicts Subjective Well-being

10.2.1 Preliminary Analysis: Using Personality Alignment to Predict the Amount Spent on Brands and Products

Prior to testing whether personality alignment (PA) predicted subjective well-being (SWB), it first had to be verified that alignment was occurring in the data using the proposed measure. To do this it was tested whether people spent more on the brands and products which aligned with their personality. This was done by investigating whether the alignment between people's personalities and the products and brands they buy predicted the amount spent on those products and brands.

Results

Hierarchical linear regression models with random intercepts were used to predict the amount spent on each brand and in each product category. Two analyses were performed: Analysis 1 predicted how much each individual spent on each brand they bought, with PA calculated using brands; and Analysis 2 predicted how much was spent in each product category a person had bought from, using PA calculated for product categories.

Both analyses followed the same format: Model 1 used overall PA to predict the amount spent, with demographics and spending as controls; Model 2 had the individual's personality and the brand/ products' personality included as additional controls. There were multiple observations per person, one for each brand/product they had purchased. The results for

Analysis 1 can be found in Table 1, and Analysis 2 in Table 2. Overall, for both products and brands, PA positively predicted spending when all other variables were controlled for. Therefore, people spent more on products and brands that aligned with their personality.

Table S1: Multiple linear regression model with random intercepts predicting the amount spent on brands using personality alignment (PA). Coefficients are *B* unstandardized.

	<i>Dependent variable:</i>	
	(log) Amount spent on different brands (£)	
	Model 1	Model 2
Overall PA (all the Big Five)	0.27*** (0.26, 0.28)	0.03*** (0.02, 0.03)
(log) Average Monthly Spend (£)	0.45*** (0.45, 0.46)	0.43*** (0.42, 0.44)
Gender: Female	0.03* (0.004, 0.05)	−0.03* (−0.05, −0.01)
Age	0.004*** (0.003, 0.004)	0.001*** (0.0003, 0.001)
Income: Less than £25,000	−0.01 (−0.04, 0.02)	−0.02 (−0.05, 0.01)
Income: £25,000 to £34,999	−0.01 (−0.06, 0.03)	−0.02 (−0.06, 0.02)
Income: £35,000 to £49,999	−0.004 (−0.02, 0.01)	−0.01 (−0.03, 0.01)
Income: £50,000 to £74,999	−0.01 (−0.03, 0.01)	−0.01 (−0.03, 0.01)
Income: £75,000 to £99,999	−0.01 (−0.02, 0.01)	−0.01 (−0.02, 0.01)
Income: £100,000 to £149,999	−0.01 (−0.04, 0.01)	−0.02 (−0.04, 0.01)
Income: £150,000 or more	−0.01 (−0.03, 0.004)	−0.02* (−0.03, −0.0005)
Person Openness		−0.01* (−0.01, −0.001)
Person Conscientiousness		0.01** (0.003, 0.01)
Person Extraversion		−0.01** (−0.01, −0.002)
Person Agreeableness		0.002 (−0.003, 0.01)
Person Neuroticism		0.01*** (0.003, 0.01)
Brand Openness		−5.01*** (−5.08, −4.94)
Brand Conscientiousness		−4.73*** (−4.86, −4.60)
Brand Extraversion		6.92*** (6.84, 7.00)
Brand Agreeableness		2.83*** (2.72, 2.95)
Brand Neuroticism		−3.06*** (−3.16, −2.95)
Observations	394,333	394,333
Log Likelihood	−710,325.00	−683,924.80
Akaike Inf. Crit.	1,420,678.00	1,367,898.00
Pseudo R Squared	0.05	0.17

p*<0.05; *p*<0.01; ***<0.001

Table 10.2: Multiple linear regression model with random intercepts predicting the amount spent on brands using personality alignment (PA). Coefficients are *B* unstandardized.

	<i>Dependent variable:</i>	
	(log) Amount spent on different products (£)	
	Model 1	Model 2
Overall PA (all the Big Five)	0.28*** (0.27, 0.29)	0.22*** (0.22, 0.23)
(log) Average Monthly Spend (£)	0.56*** (0.55, 0.57)	0.55*** (0.54, 0.56)
Gender: Female	-0.05*** (-0.07, -0.03)	-0.07*** (-0.09, -0.05)
Age	0.001** (0.0002, 0.001)	0.0001 (-0.0002, 0.0005)
Income: Less than £25,000	0.03 (-0.002, 0.06)	0.03* (0.004, 0.06)
Income: £25,000 to £34,999	0.04* (0.002, 0.08)	0.05** (0.01, 0.10)
Income: £35,000 to £49,999	-0.02* (-0.03, -0.0005)	-0.02* (-0.04, -0.004)
Income: £50,000 to £74,999	-0.01 (-0.03, 0.004)	-0.02* (-0.03, -0.001)
Income: £75,000 to £99,999	-0.001 (-0.02, 0.02)	-0.001 (-0.02, 0.02)
Income: £100,000 to £149,999	0.01 (-0.02, 0.03)	0.01 (-0.01, 0.03)
Income: £150,000 or more	-0.01 (-0.03, 0.005)	-0.01 (-0.03, 0.001)
Person Openness		-0.002 (-0.01, 0.003)
Person Conscientiousness		-0.003 (-0.01, 0.001)
Person Extraversion		-0.003 (-0.01, 0.0004)
Person Agreeableness		-0.002 (-0.01, 0.003)
Person Neuroticism		0.01*** (0.01, 0.01)
Product Openness		-4.05*** (-4.12, -3.98)
Product Conscientiousness		1.01*** (0.92, 1.11)
Product Extraversion		5.24*** (5.16, 5.32)
Product Agreeableness		-0.13* (-0.25, -0.01)
Product Neuroticism		1.07*** (0.97, 1.17)
Observations	488,153	488,153
Log Likelihood	-873,465.30	-863,485.40
Akaike Inf. Crit.	1,746,959.00	1,727,019.00
Psuedo R Squared	0.07	0.11

*p<0.05; **p<0.01; ***p<0.001

10.2.2 Tables of Brand and Product Descriptives

Table 10.3: The 100 most popular brands used in Analysis 1. Brands are ordered by their absolute mean Z-score (Abs. Mean) across their expected trait values for the Big Five. A higher absolute mean Z-score signals that the brand has a more extreme indicated personality across the Big Five trait dimensions. Mean var is the mean variance across the Big Five. O (openness to new experiences), C (conscientiousness), E (extraversion), A (agreeableness), and ES (emotional stability) are the Z-scores of the expected trait score for the brand. Sales is the number of times that brand was purchased by customers in the sample.

Brand	Abs. Mean	Mean Var	O	C	E	A	ES	Sales
BODYFORM	2.37	1.63	-1.61	-1.92	-3.67	-2.67	-1.96	10,630
IMPULSE	2.24	1.65	-2.30	-1.83	-2.79	-1.94	-2.35	9,889
CLARINS	2.11	1.69	1.33	1.96	2.96	2.17	2.11	12,967
CAREX	1.83	1.67	-1.84	-1.76	-2.14	-1.87	-1.52	9,870
NATURAL COLLECTION	1.80	1.68	-1.96	-1.44	-2.14	-1.42	-2.04	22,881
17	1.56	1.68	-0.62	-2.08	-1.20	-1.49	-2.43	28,775
NUTRITIOUS	1.50	1.68	2.95	1.58	0.08	0.98	1.90	25,609
WRIGLEYS	1.45	1.67	0.04	-2.38	-0.77	-1.76	-2.32	12,977
INNOCENT	1.42	1.68	1.79	1.33	-1.41	0.37	2.23	16,243
HEINZ	1.37	1.60	-0.50	-1.46	-0.26	-3.02	-1.63	10,138
OLAY	1.37	1.65	0.22	1.95	0.48	2.33	1.88	12,204
BARRY M	1.33	1.72	0.88	-1.97	0.02	-1.63	-2.15	15,809
ESSENTIALS	1.31	1.66	-1.84	-1.48	-0.89	-1.47	-0.88	48,933
ELLAS KITCHEN	1.28	1.56	2.37	1.08	1.87	0.86	0.24	11,030
LUCOZADE	1.20	1.69	-0.50	-1.08	-1.32	-1.42	-1.67	10,805
LEE STAFFORD	1.15	1.70	-0.49	-1.68	-1.45	-0.83	-1.29	9,729
TAMPAX	1.14	1.67	1.97	1.15	1.43	0.77	0.35	10,840
NAKED	1.13	1.71	2.04	0.20	-0.98	-0.95	1.46	12,399
PEPSI	1.12	1.66	0.47	-2.40	-1.02	-0.96	-0.75	10,006
CLINIQUE	1.11	1.67	-0.74	1.62	0.50	1.26	1.44	29,494
BOOTS-HOSIERY	1.08	1.67	1.34	0.56	1.45	1.17	0.90	14,736
SMILE	1.07	1.67	0.27	1.38	0.47	1.73	1.52	25,112
CHAMPNEYS	1.07	1.64	0.32	1.02	0.89	1.55	1.56	18,433
BOOTS SKINCARE	1.07	1.64	-0.60	1.13	1.00	1.37	1.23	105,043
AUSSIE	1.03	1.66	1.00	0.93	1.34	0.90	1.00	27,161
ACCESSORY CROSSING	1.01	1.68	-0.95	-2.34	-0.04	-0.26	-1.45	9,741
BOOTS-MBU OTHER	0.98	1.66	-1.21	-0.52	-1.61	-0.49	-1.08	43,879
COLLECTION 2000	0.98	1.70	0.11	-1.68	-0.26	-1.19	-1.64	13,163
HIPP	0.95	1.60	1.07	1.12	1.32	-1.08	0.18	11,147
SHAPERS	0.94	1.67	1.77	0.71	0.78	1.22	0.24	128,183
CADBURY'S	0.94	1.69	-0.99	-0.66	-1.18	-0.55	-1.29	29,664
BOURJOIS	0.90	1.74	1.35	-1.31	0.43	-0.29	-1.10	15,480
COCA COLA	0.86	1.69	1.36	0.04	1.64	1.10	-0.18	37,212
PANTENE	0.82	1.66	-0.20	0.81	-0.66	1.42	0.99	10,388
BOOTS CAT MAN	0.81	1.67	1.54	0.71	0.31	0.76	0.73	157,428
CENTRAL POSTINGS	0.80	1.69	0.82	1.14	0.38	1.23	0.46	26,547
REVLON	0.79	1.71	0.20	-1.27	-0.14	-1.07	-1.28	15,122
SANCTUARY	0.78	1.64	0.16	0.97	0.99	0.93	0.87	38,023
SENSODYNE	0.78	1.65	-1.14	-0.57	-0.94	-0.98	-0.24	16,676
COW AND GATE	0.77	1.58	0.56	0.18	1.04	-0.90	-1.17	14,923
BATISTE	0.77	1.65	0.87	-0.21	1.02	-0.81	-0.94	14,984
SOLTAN	0.76	1.65	-0.15	1.17	0.81	0.64	1.03	27,969
BOOTS OTHER	0.75	1.65	0.18	0.83	1.00	1.02	0.74	24,058
DELICIOUS	0.75	1.66	1.36	0.78	0.06	0.62	0.94	121,760
PAMPERS	0.73	1.60	-1.03	-0.34	0.95	-0.83	-0.49	12,510

Brand	Abs. Mean	Mean Var	O	C	E	A	ES	Sales
BOOTS PHARMA.	0.71	1.67	0.18	0.70	0.46	1.18	1.02	329,035
VALUE HEALTH	0.70	1.67	0.64	1.00	0.47	0.91	0.48	26,497
BRECON CARREG	0.70	1.66	1.09	0.57	0.65	0.92	0.26	25,947
BOOTS-BABY FEED	0.69	1.61	0.35	-0.19	0.88	-1.52	-0.49	9,897
P&G	0.68	1.69	-0.50	-1.07	-0.84	-0.51	-0.46	11,663
BENEFIT	0.67	1.68	0.80	-0.58	1.60	-0.14	0.24	12,753
MAYBELLINE	0.67	1.71	0.80	-0.58	0.87	-0.44	-0.66	33,022
LYNX	0.66	1.63	-1.19	0.46	-1.21	-0.17	0.28	21,179
EVIAN	0.66	1.67	1.48	0.75	0.44	0.14	0.49	10,626
WILKINSON SWORD	0.63	1.69	0.57	1.25	0.23	0.84	0.28	10,200
TRESEMME	0.61	1.64	-0.92	-0.60	-0.35	-0.64	-0.56	20,914
CHARLESWORTHINGTON	0.60	1.66	0.07	0.83	0.41	1.12	0.57	17,560
JOHNSON & JOHNSON	0.58	1.66	-0.99	-0.59	0.20	-0.76	-0.34	16,391
SANEX	0.57	1.63	-0.93	-0.18	-1.34	-0.34	-0.08	11,565
RIMMEL	0.56	1.67	0.19	-0.68	0.30	-0.73	-0.92	49,270
ELNETT	0.56	1.67	-0.64	0.33	0.17	0.84	0.84	12,441
ALWAYS	0.56	1.65	0.45	0.14	-0.64	0.95	-0.64	23,870
RADOX	0.54	1.63	-1.17	0.34	-0.63	-0.34	0.24	27,208
HERBAL ESSENCES	0.54	1.65	0.70	0.52	0.38	0.79	0.32	12,842
OGX	0.52	1.66	1.23	0.09	0.96	-0.08	-0.24	13,531
BOOTS-PHOTO	0.51	1.64	-1.04	0.19	-0.36	0.45	0.52	21,321
NUROFEN	0.51	1.67	-0.85	-0.30	-0.60	0.76	0.05	13,674
LISTERINE	0.50	1.69	0.20	0.31	0.51	0.43	1.06	16,644
ESTEE LAUDER	0.50	1.71	-0.50	-0.49	1.20	0.19	0.12	10,583
IMPERIAL LEATHER	0.50	1.68	-0.89	-0.56	-0.44	-0.57	-0.02	12,517
EXPERT	0.50	1.64	-0.59	0.32	-0.58	0.21	0.78	18,383
BB HEALTHCARE	0.49	1.72	-0.65	-0.08	-1.05	0.65	-0.02	16,564
DVD	0.47	1.66	-0.80	-0.50	0.39	-0.58	0.07	21,053
VASELINE IC	0.46	1.65	0.29	0.70	0.52	0.50	0.29	13,383
WALKERS	0.44	1.66	0.41	0.58	-0.38	0.47	0.35	52,674
SIMPLE	0.43	1.67	-0.78	0.65	-0.10	0.38	0.27	30,370
GILLETTE	0.42	1.65	-0.33	0.48	0.25	-0.36	0.70	31,874
NEUTROGENA	0.42	1.70	0.24	0.99	-0.33	0.32	0.20	11,222
JOHN FRIEDA	0.41	1.68	-0.37	0.08	0.39	0.58	0.63	40,978
ORIGINAL SOURCE	0.41	1.66	0.54	0.45	0.34	-0.37	0.32	12,408
SALLY HANSEN	0.40	1.69	-0.57	-0.58	0.41	-0.40	0.07	11,085
KLEENEX	0.40	1.69	-0.46	-0.17	-1.00	-0.29	-0.06	13,872
MOTHERCARE	0.39	1.58	-0.16	0.21	0.91	-0.01	0.68	34,236
ESSENTIALS ACCESSO	0.38	1.68	-0.88	-0.44	-0.23	-0.19	-0.15	10,093
COLGATE	0.37	1.67	-0.49	0.17	-0.51	-0.39	0.31	50,906
S&G INDULGENT BATH	0.37	1.67	0.37	0.37	-0.42	0.69	-0.001	27,171
FCUK	0.37	1.63	-0.81	0.32	-0.12	0.39	0.20	10,812
BOOTS COSMETICS	0.36	1.66	-1.05	-0.35	0.18	0.04	0.18	121,257
MAX FACTOR	0.36	1.69	-0.41	-0.66	0.37	0.25	-0.11	32,876
GARNIER	0.36	1.68	0.62	0.37	0.19	0.31	-0.31	31,266
BOTANICS	0.35	1.68	0.72	0.31	-0.30	0.04	-0.40	23,990
DOVE	0.33	1.64	-0.60	0.45	-0.33	0.17	0.10	56,190
TED BAKER	0.32	1.64	-0.46	0.14	-0.05	0.52	0.42	14,760
L'OREAL	0.29	1.67	0.20	-0.52	0.27	-0.34	-0.15	95,822
NIVEA	0.29	1.66	-0.71	0.33	-0.09	-0.16	0.16	75,759
ORAL B	0.29	1.67	-0.33	0.36	-0.08	0.02	0.63	38,828
SURE	0.28	1.66	-0.52	0.24	0.01	-0.35	0.31	37,847
VO5	0.26	1.66	-0.64	-0.10	-0.27	-0.25	0.03	13,758

Table 10.4: The 100 most popular product categories used in analysis 2. Products are ordered by their absolute mean Z-score (Abs. Mean) across their expected trait values for the Big Five. A higher absolute mean Z-score signals that the product has a more extreme indicated personality across the Big Five trait dimensions. Mean Var is the mean variance across the Big Five. O (openness to new experiences) , C (conscientiousness), E (extraversion), A (agreeableness), and ES (emotional stability) are the Z-scores of the expected trait score for the product category. Sales is the number of times that product was purchased by customers in the sample.

Product Category	Abs. Mean	Mean Var	O	C	E	A	ES	Sales
Smoking Cessation	3.44	1.66	2.68	1.56	4.07	4.06	4.82	15,489
Baby Toiletries	2.51	1.66	-2.40	-2.91	-1.34	-3.53	-2.38	24,167
Contact Lens Care	2.27	1.59	2.41	3.24	2.14	1.53	2.03	13,458
Tights	1.96	1.68	2.74	1.18	2.70	1.87	1.30	11,100
Sanitary Towels	1.90	1.66	-1.21	-1.53	-2.81	-1.48	-2.48	31,395
Salads	1.59	1.69	3.25	0.75	1.64	0.90	1.42	33,405
Incontinence Pads	1.54	1.70	-1.53	-2.38	-2.41	0.88	-0.47	14,995
Decongestants	1.50	1.65	0.47	1.57	1.37	2.21	1.91	22,182
Hand Wash	1.38	1.67	-1.65	-0.76	-1.78	-1.30	-1.39	33,801
Sweet Snacks	1.37	1.67	-0.41	-1.98	-1.62	-1.37	-1.47	141,068
Dietary	1.27	1.72	-0.14	-2.10	-2.00	-0.67	-1.43	11,245
Body Sprays	1.23	1.68	-1.41	-0.60	-1.93	-1.16	-1.05	31,089
Jewellery	1.22	1.68	-0.95	-2.67	-0.06	-0.68	-1.74	10,536
First Aid Accs	1.22	1.67	1.46	1.37	0.62	2.01	0.64	10,513
Nappies	1.21	1.56	-1.39	-1.25	-0.02	-2.31	-1.08	21,578
Newspapers & Mags	1.20	1.66	-0.97	-0.85	-1.69	-0.96	-1.51	63,744
Fragrance Gift	1.11	1.64	-1.45	-1.28	-0.73	-1.24	-0.83	12,995
Baby Wipes	1.10	1.63	-0.98	-1.55	-0.02	-1.73	-1.23	23,151
Formula/Milk	1.06	1.58	-0.36	-1.03	0.19	-1.91	-1.83	14,286
Denture Products	1.03	1.63	-0.82	0.14	-1.35	1.35	1.50	12,133
Chilled Drinks	0.98	1.70	1.95	0.49	-0.95	0.31	1.21	21,039
Fresh Fruit	0.96	1.67	2.12	0.69	-0.83	0.30	0.88	20,545
Foot Skincare	0.89	1.64	-0.79	1.32	0.13	0.79	1.44	11,568
Grooming Tools	0.88	1.70	1.07	0.62	1.07	0.99	0.68	13,201
Depilatories	0.88	1.69	1.39	0.61	1.45	0.44	0.50	15,939
Baby Wet Foods	0.85	1.58	0.96	0.37	1.00	-1.34	-0.59	40,114
Eye Make Up Remover	0.79	1.71	-0.88	-0.82	-0.16	-1.15	-0.91	13,628
Cotton Wool	0.79	1.67	0.42	1.39	0.68	0.92	0.52	41,505
Cosmetic Proms	0.77	1.67	0.66	0.17	1.12	1.25	0.64	14,202
Self Tan Products	0.76	1.66	0.85	-0.48	1.71	-0.64	-0.10	14,849
Weight Management	0.76	1.70	0.52	-1.61	0.03	-0.43	-1.19	16,865
Eye Cosmetics	0.75	1.67	-0.20	-1.43	0.17	-1.19	-0.77	141,794
Tampons	0.75	1.68	1.34	1.08	1.03	0.05	0.25	21,397
Sandwiches	0.75	1.66	1.24	0.99	-0.16	0.59	0.74	124,491
Greeting Cards	0.74	1.60	-0.48	0.57	0.49	1.08	1.06	12,686
Sunscreen/Sunblock	0.74	1.65	-0.17	1.09	1.00	0.37	1.05	41,586
Throat Remedies	0.73	1.65	-1.14	-1.02	-0.96	-0.49	-0.03	27,901
Hair Colours	0.71	1.69	1.88	-0.50	0.79	0.36	-0.04	45,286
Facial Scrubs	0.69	1.68	0.06	-0.31	-0.76	-1.14	-1.19	16,150
Baby Snacks	0.68	1.64	1.32	0.44	1.48	0.08	-0.09	18,034
Nail Care Products	0.68	1.67	-0.90	-0.97	-0.55	-0.70	-0.29	18,819
Cold/Flu Remedies	0.66	1.67	0.23	0.88	0.81	0.47	0.91	24,359
Feminine Healthcare	0.66	1.69	1.09	0.57	0.32	0.69	-0.63	10,500
Panty Liners	0.66	1.62	-0.69	0.17	-1.29	-0.46	-0.68	26,854
Shaving Preparations	0.66	1.65	-0.20	1.40	0.47	0.03	1.17	19,005
ROI Bag Levy	0.65	1.68	0.33	-1.21	0.07	-0.34	-1.33	30,380
Nail Enamel Removers	0.61	1.67	-0.63	-0.89	0.45	-0.96	0.11	19,419
Nail Enamels	0.61	1.67	0.04	-0.97	-0.24	-0.55	-1.23	82,575
DVDs Pre-Recorded	0.61	1.66	-0.83	-0.79	0.29	-1.05	-0.08	21,071
Desserts	0.59	1.69	1.68	-0.13	-0.39	-0.68	-0.08	27,788

Product Category	Abs. Mean	Mean Var	O	C	E	A	ES	Sales
Cosmetic Applicators	0.59	1.69	0.58	-0.51	0.49	-0.59	-0.75	18,782
Cough Remedies	0.59	1.66	-1.22	-0.21	-0.82	0.14	0.55	20,140
Facial Tissues	0.58	1.67	-0.47	-0.78	-0.81	-0.51	-0.34	52,473
Face Cosmetics	0.58	1.67	-0.40	-0.58	-0.24	-0.71	-0.95	138,054
Savoury Snacks	0.57	1.67	1.10	0.77	0.13	0.76	0.11	98,457
Health Foot Care Acc	0.56	1.67	-0.41	0.82	0.35	0.62	0.59	11,727
Wash Accessories	0.53	1.67	0.03	0.99	0.43	0.45	0.75	22,586
Laxatives	0.52	1.71	-0.17	0.72	0.46	0.58	-0.65	13,569
Plasters	0.51	1.67	0.23	0.67	0.06	0.89	0.71	17,138
Bandages & Dressings	0.50	1.68	-0.02	0.24	-0.03	1.20	1.05	13,056
Eye Preps	0.50	1.64	0.15	1.11	0.04	1.00	0.22	23,966
Damaged Skin	0.50	1.66	-0.57	-0.45	-0.85	0.14	-0.51	16,103
Dental Floss/Tape	0.48	1.66	-0.09	0.83	-0.82	0.26	0.38	13,597
Nail Care Access.	0.46	1.67	0.05	0.46	0.96	0.36	0.48	17,951
Indigestion Remedies	0.46	1.67	-0.10	-0.36	0.16	1.18	0.48	26,479
Childrens P/Relief	0.45	1.61	-0.26	0.56	0.82	-0.41	-0.21	12,502
Pain Relief Oral	0.45	1.66	-0.69	-0.46	-0.40	0.40	-0.29	121,415
Drinks	0.45	1.67	1.12	-0.26	-0.22	-0.13	-0.52	203,909
Lip Cosmetics	0.45	1.67	0.25	-0.91	0.40	-0.16	-0.51	73,154
Facial Moisturisers	0.44	1.66	-0.45	0.47	0.29	0.52	0.47	194,441
Vitamins	0.44	1.67	-0.10	0.91	-0.27	0.46	0.45	96,602
Body Moisturisers	0.43	1.65	-0.42	0.94	-0.09	0.40	0.32	91,254
EDTs	0.42	1.65	-0.36	0.23	0.59	0.38	0.55	27,631
Complemen. Medicines	0.42	1.68	0.20	0.69	0.17	1.03	-0.02	29,235
Cosmetic Gift	0.41	1.65	-0.39	-0.46	-0.43	-0.29	-0.50	32,828
Anti-Perspirants	0.41	1.66	-0.31	1.04	-0.19	0.16	0.35	129,121
Hair Tools & Accs	0.41	1.69	0.23	-0.61	-0.17	-0.46	-0.58	24,640
Childrens Clothing	0.41	1.58	-0.11	0.17	0.91	-0.23	0.63	34,232
Toiletry Gift	0.40	1.63	-0.86	0.23	-0.63	0.06	-0.24	92,466
Facial Cleansers	0.40	1.67	-0.04	-0.37	-0.39	-0.80	-0.42	52,388
Bath Additives	0.38	1.66	-0.20	0.52	-0.67	-0.14	0.37	23,838
EDPs	0.37	1.66	-0.52	-0.22	0.72	-0.23	0.18	16,429
Dental Accessories	0.37	1.65	-0.50	0.33	-0.36	-0.10	0.56	16,885
Food Gift	0.36	1.62	0.06	0.38	-0.45	0.21	-0.69	10,554
Shower Gels	0.35	1.65	-0.62	0.55	-0.40	-0.10	0.09	100,948
Hair Styling	0.34	1.66	-0.62	-0.43	-0.11	-0.39	-0.15	99,404
Hair Combs/Brushes	0.33	1.66	-0.54	-0.26	-0.56	-0.28	0.01	16,039
First Time DP	0.31	1.62	-1.20	-0.06	-0.13	0.13	0.04	21,155
Mouth Fresheners	0.31	1.69	-0.24	0.27	-0.21	0.002	0.81	45,566
Eye Gels/Creams	0.30	1.67	-0.03	0.36	0.49	0.54	0.09	16,216
Razors & Blades	0.30	1.67	-0.24	0.65	-0.04	-0.47	-0.09	37,129
Toothpaste	0.28	1.66	-0.17	0.18	-0.48	-0.14	0.41	84,168
Lipsalves/Balms	0.27	1.66	-0.11	0.69	-0.20	-0.32	0.004	24,456
Hayfever Relief	0.27	1.64	-0.13	0.15	0.20	0.52	0.33	27,122
Hair Conditioner	0.26	1.66	-0.14	-0.09	-0.59	-0.23	-0.27	101,479
Topical Pain Relief	0.26	1.66	-0.66	0.01	-0.06	0.40	-0.17	17,606
Diarrhoea Remedies	0.23	1.66	0.22	-0.34	-0.22	0.25	0.11	13,148
Hair Shampoo	0.22	1.66	-0.22	0.26	-0.35	-0.11	-0.18	126,588
Toothbrushes	0.21	1.68	0.05	0.51	0.16	-0.12	0.23	34,087
Skin Care Wipes	0.14	1.67	-0.31	-0.02	0.12	-0.18	0.07	42,657

10.2.3 Personality Distributions

Personality of participants in the sample for Analysis 1 (brands)

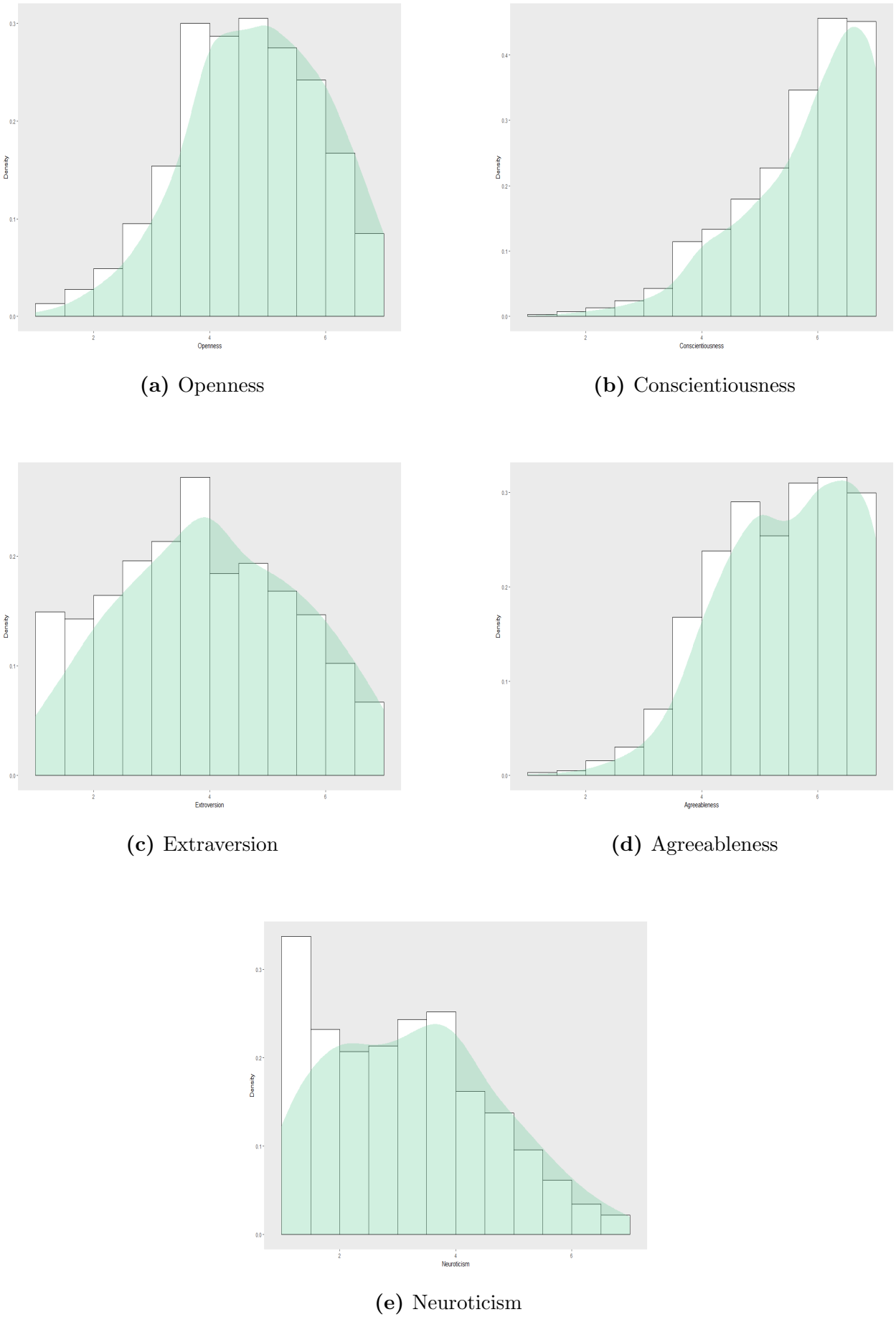


Figure 10.1: Distribution of participant’s Big Five personality traits in the brand sample (those who had bought at least one of the 100 most popular brands used in the study). Traits were measured on a likert scale from 1-7.

10.2.4 Expected Brand Personality

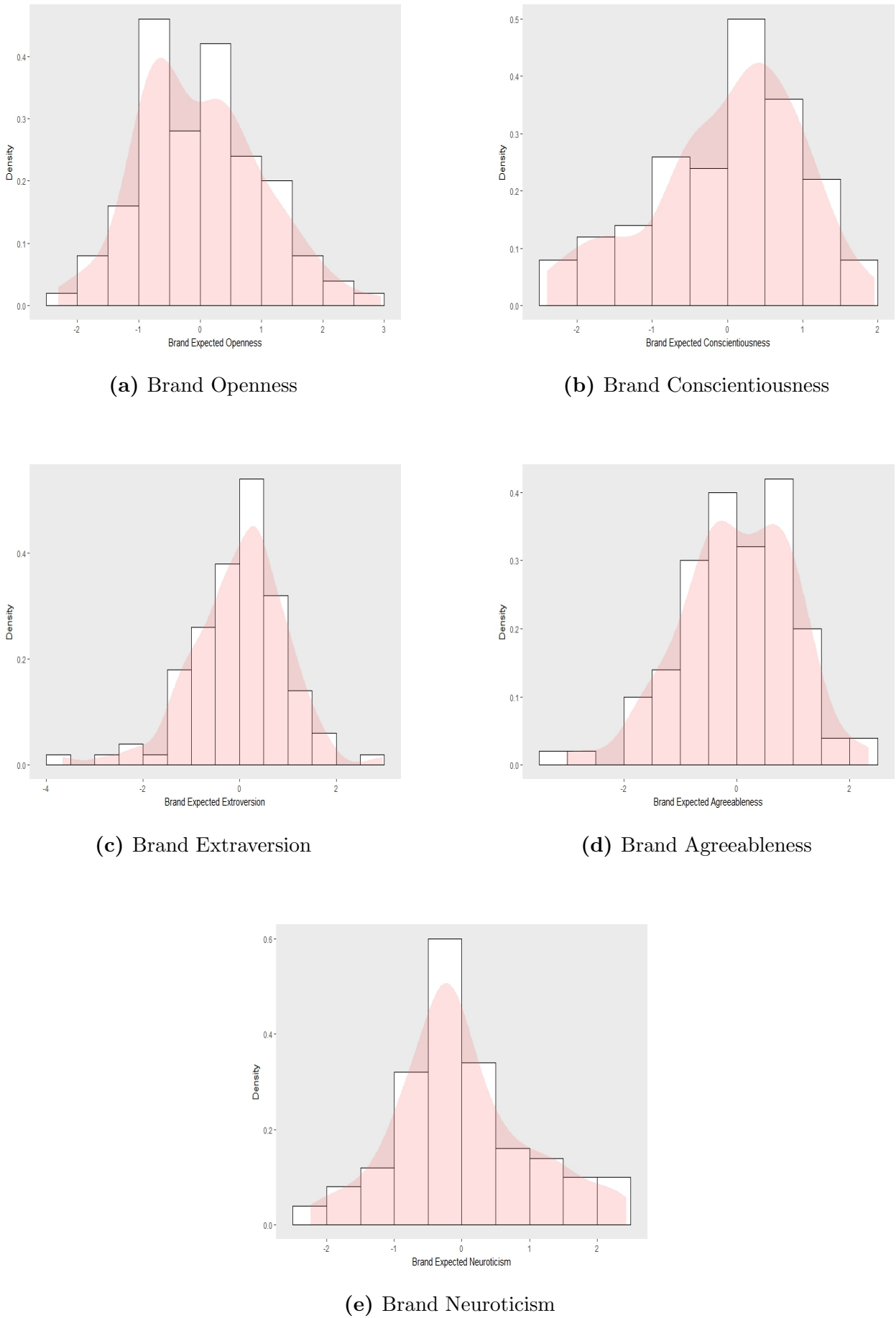


Figure 10.2: Distribution of the z-standardised expected brand personalities

10.2.5 Personality of Participants in the Sample for Analysis 2 (Products)

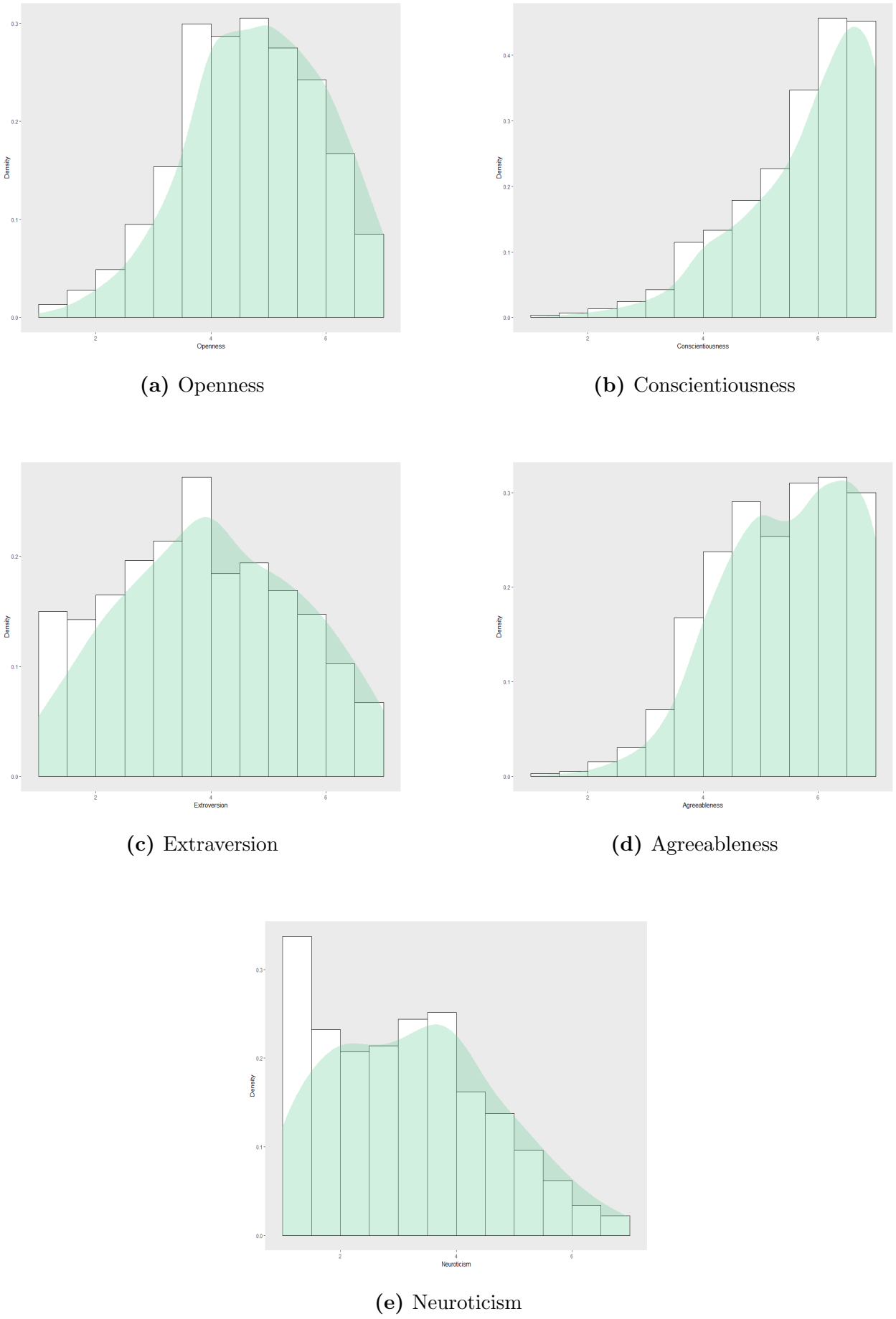


Figure 10.3: Distribution of participant’s Big Five personality traits in the product sample (those who had bought at least one of the 100 most popular product categories used in the study). Traits were measured on a likert scale from 1-7.

10.2.6 Expected Product Personality

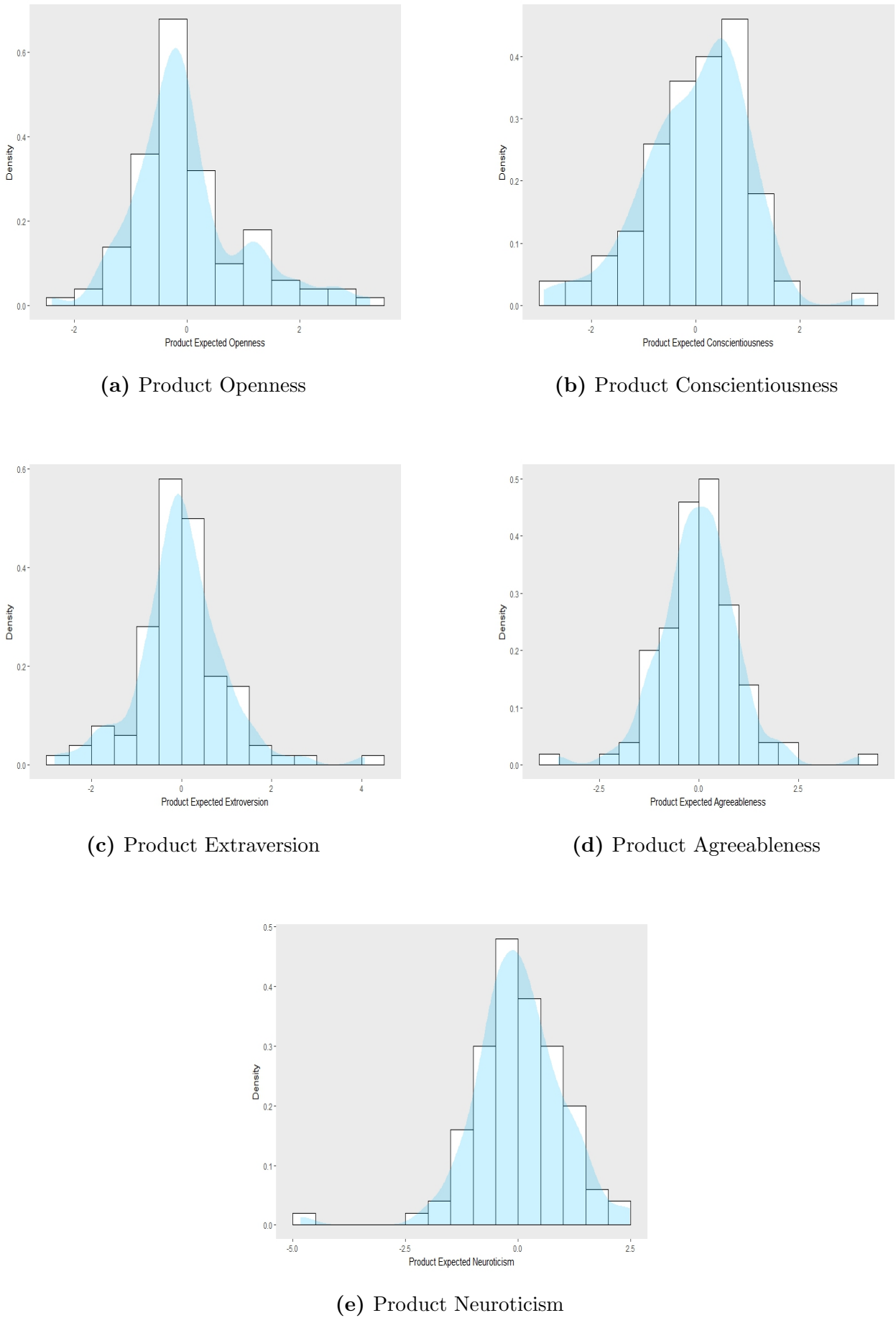


Figure 10.4: Distribution of the z-standardised expected product personalities

10.2.7 Correlations

Table 10.5: Brand analysis: Pearson correlation coefficients. For gender, Male = 1; ‘Avg.£’ = Average monthly spend; SWB = subjective well-being; PA = Personality Alignment; for income the reference class is ‘Prefer not to say’; I1 = Income: ‘Less than £25,000’; I2 = Income: ‘£25,000-34,999’; I3 = Income: ‘£35,000-49,999’; I4 = Income: ‘£50,000-74,999’; I5 = Income: ‘£75,000-94,999’; I6 = Income: ‘£100,000-149,999’; I7 = ‘£150,000 or more’

	Age	Gender	Avg.£	I1	I2	I3	I4	I5	I6	I7	O	C	E	A	N	SWB	PA ^O	PA ^C	PA ^E	PA ^A	PA ^N
Age	1	0.13	-0.04	0.12	0.03	-0.08	-0.12	-0.09	-0.06	-0.02	-0.01	0.12	0.05	0.21	-0.20	0.17	0.02	0.09	0.06	0.08	0.10
Gender	0.13	1	-0.10	0	0.01	0	0.01	0.02	0.03	0.01	-0.02	-0.01	-0.06	-0.09	-0.07	-0.02	-0.01	0.03	0.02	-0.03	0.06
Avg.£	-0.04	-0.10	1	-0.14	-0.03	0.02	0.08	0.09	0.06	0.08	0.02	0.02	0.07	0.02	-0.02	0.05	0.01	0	-0.03	0	-0.02
I1	0.12	0	-0.14	1	-0.24	-0.25	-0.22	-0.13	-0.09	-0.06	-0.01	-0.06	-0.07	0.05	0.07	-0.13	-0.02	-0.05	-0.01	0	-0.05
I2	0.03	0.01	-0.03	-0.24	1	-0.20	-0.18	-0.11	-0.08	-0.05	0.01	-0.02	0.01	0.01	-0.01	0	0.01	0	0.01	0	0.01
I3	-0.08	0	0.02	-0.25	-0.20	1	-0.18	-0.11	-0.08	-0.05	-0.01	-0.01	0.01	-0.03	0	0.03	0.01	0	0	0	0.01
I4	-0.12	0.01	0.08	-0.22	-0.18	-0.18	1	-0.10	-0.07	-0.05	0	0.02	0.06	-0.04	-0.02	0.03	0	0.02	-0.02	0	0
I5	-0.09	0.02	0.09	-0.13	-0.11	-0.11	-0.10	1	-0.04	-0.03	0.02	0.03	0.04	-0.02	-0.03	0.03	-0.01	0.01	-0.01	0	0.02
I6	-0.06	0.03	0.06	-0.09	-0.08	-0.08	-0.07	-0.04	1	-0.02	0.01	0.02	0.02	-0.03	-0.02	0.02	0.01	0.01	-0.01	-0.02	0
I7	-0.02	0.01	0.08	-0.06	-0.05	-0.05	-0.05	-0.03	-0.02	1	0.03	0.03	0.03	0	-0.04	0.03	-0.01	0	0	-0.01	-0.02
O	-0.01	-0.02	0.02	-0.01	0.01	-0.01	0	0.02	0.01	0.03	1	0.06	0.36	0.15	-0.21	0.09	0.09	-0.01	0.02	0.04	0.01
C	0.12	-0.01	0.02	-0.06	-0.02	-0.01	0.02	0.03	0.06	1	0.07	0.21	0.21	-0.27	-0.27	0.20	-0.03	0.67	-0.02	0.03	0.06
E	0.05	-0.06	0.07	-0.07	0.01	0.01	0.06	0.04	0.02	0.03	0.36	0.07	1	0.06	-0.23	0.20	0.01	0.02	0.13	0.03	0.05
A	0.21	-0.09	0.02	0.05	0.01	-0.03	-0.04	-0.02	-0.03	0	0.15	0.21	0.06	1	-0.29	0.16	-0.01	0.11	0	0.38	0.04
N	-0.20	-0.07	-0.02	0.07	-0.01	0	-0.02	-0.03	-0.02	-0.04	-0.21	-0.27	-0.23	-0.29	1	-0.40	-0.04	-0.17	-0.07	-0.10	-0.29
SWB	0.17	-0.02	0.05	-0.13	0	0.03	0.03	0.03	0.02	0.03	0.09	0.20	0.20	0.16	-0.40	1	0.02	0.13	0.07	0.09	0.18
PA ^O	0.02	-0.01	0.01	-0.02	0.01	0.01	0	-0.01	0.01	-0.01	0.09	-0.03	0.01	-0.01	-0.04	0.02	1	0.05	0.19	0.08	0.14
PA ^C	0.09	0.03	0	-0.05	0	0	0.02	0.01	0.01	0	-0.01	0.67	0.02	0.11	-0.17	0.13	0.05	1	0.04	0.11	0.18
PA ^E	0.06	0.02	-0.03	-0.01	0.01	0	-0.02	-0.01	-0.01	0	0.02	-0.02	0.13	0	-0.07	0.07	0.19	0.04	1	0.07	0.16
PA ^A	0.08	-0.03	0	0	0	0	0	0	-0.02	-0.01	0.04	0.03	0.03	0.38	-0.10	0.09	0.08	0.11	0.07	1	0.16
PA ^N	0.10	0.06	-0.02	-0.05	0.01	0.01	0	0.02	0	-0.02	0.01	0.06	0.05	0.04	-0.29	0.18	0.14	0.18	0.16	0.16	1

Table 10.6: Product analysis: Pearson correlation coefficients. For gender, Male = 1; ‘Avg. £’ = Average monthly spend; SWB = subjective well-being; PA = Personality Alignment; for income the reference class is ‘Prefer not to say’; I1 = Income: ‘Less than £25,000’; I2 = Income: ‘£25,000-34,999’; I3 = Income: ‘£35,000-49,999’; I4 = Income: ‘£50,000-74,999’; I5 = Income: ‘£75,000-94,999’; I6 = Income: ‘£100,000-149,999’; I7 = ‘£150,000 or more’

	Age	Gender	Avg.£	I1	I2	I3	I4	I5	I6	I7	O	C	E	A	N	SWB	PA ^O	PA ^C	PA ^E	PA ^A	PA ^N
Age	1	0.12	-0.04	0.13	0.03	-0.08	-0.12	-0.09	-0.06	-0.02	-0.02	0.12	0.03	0.22	-0.19	0.15	0.03	0.06	0.05	0.04	0.08
Gender	0.12	1	-0.10	0	0.01	0	0.01	0.02	0.03	0.01	-0.02	-0.01	-0.05	-0.09	-0.07	-0.02	-0.01	0.01	0.01	-0.05	0.02
Avg.£	-0.04	-0.10	1	-0.14	-0.03	0.02	0.08	0.08	0.06	0.08	0.02	0.02	0.07	0.02	-0.02	0.04	0	-0.03	-0.01	-0.03	
I1	0.13	0	-0.14	1	-0.25	-0.25	-0.22	-0.13	-0.09	-0.06	-0.02	-0.06	-0.07	0.05	0.07	-0.13	-0.01	-0.05	0	0	-0.03
I2	0.03	0.01	-0.03	-0.25	1	-0.20	-0.18	-0.11	-0.08	-0.05	0.02	-0.01	0.01	0.01	-0.01	0.01	0	0.01	0.01	0.01	0.01
I3	-0.08	0	0.02	-0.25	-0.20	1	-0.18	-0.11	-0.08	-0.05	-0.01	-0.01	0.01	-0.03	0	0.03	0.01	0	0	0.01	0.01
I4	-0.12	0.01	0.08	-0.22	-0.18	-0.18	1	-0.10	-0.07	-0.05	0	0.02	0.06	-0.04	-0.02	0.03	-0.01	0.02	-0.03	0	-0.01
I5	-0.09	0.02	0.08	-0.13	-0.11	-0.11	-0.10	1	-0.04	-0.03	0.02	0.03	0.04	-0.02	-0.03	0.04	-0.01	0.01	-0.01	0.01	0.01
I6	-0.06	0.03	0.06	-0.09	-0.08	-0.08	-0.07	-0.04	1	-0.02	0.01	0.02	0.03	-0.03	-0.02	0.02	0.01	0.01	-0.01	-0.03	-0.01
I7	-0.02	0.01	0.08	-0.06	-0.05	-0.05	-0.05	-0.03	-0.02	1	0.03	0.03	0.03	0	-0.04	0.03	-0.02	0	-0.01	-0.01	-0.02
O	-0.02	-0.02	0.02	-0.02	0.02	-0.01	0	0.02	0.01	0.03	1	0.07	0.36	0.14	-0.21	0.10	0.03	-0.01	-0.01	0.04	0
C	0.12	-0.01	0.02	-0.06	-0.01	-0.01	0.02	0.03	0.02	0.03	0.07	1	0.08	0.21	-0.27	0.20	-0.03	0.65	-0.03	0.03	0.04
E	0.03	-0.05	0.07	-0.07	0.01	0.01	0.06	0.04	0.03	0.03	0.36	0.08	1	0.05	-0.23	0.20	-0.01	0.02	0.05	0.02	0.03
A	0.22	-0.09	0.02	0.05	0.01	-0.03	-0.04	-0.02	-0.03	0	0.14	0.21	0.05	1	-0.29	0.15	-0.02	0.10	-0.01	0.33	0.02
N	-0.19	-0.07	-0.02	0.07	-0.01	0	-0.02	-0.03	-0.02	-0.04	-0.21	-0.27	-0.23	-0.29	1	-0.40	-0.02	-0.15	-0.04	-0.08	-0.22
SWB	0.15	-0.02	0.04	-0.13	0.01	0.03	0.03	0.04	0.02	0.03	0.10	0.20	0.20	0.15	-0.40	1	0	0.12	0.05	0.07	0.14
PA ^O	0.03	-0.01	0	-0.01	0	0.01	-0.01	-0.01	0.01	-0.02	0.03	-0.03	-0.01	-0.02	-0.02	0	1	0.05	0.20	0.08	0.14
PA ^C	0.06	0.01	0	-0.05	0.01	0	0.02	0.01	0.01	0	-0.01	0.65	0.02	0.10	-0.15	0.12	0.05	1	0.04	0.10	0.15
PA ^E	0.05	0.01	-0.03	0	0.01	0	-0.03	-0.01	-0.01	-0.01	-0.01	-0.03	0.05	-0.01	-0.04	0.05	0.20	0.04	1	0.07	0.16
PA ^A	0.04	-0.05	-0.01	0	0.01	0.01	0	0.01	-0.03	-0.01	0.04	0.03	0.02	0.33	-0.08	0.07	0.08	0.10	0.07	1	0.15
PA ^N	0.08	0.02	-0.03	-0.03	0.01	0.01	-0.01	0.01	-0.01	-0.02	0	0.04	0.03	0.02	-0.22	0.14	0.14	0.15	0.16	0.15	1

10.3 Chapter 6: Understanding the Predictors of Plastic Bag Consumption in Big Data, Analysis 1

10.3.1 Analysis 1: Meta-Parameters

The meta parameter values considered in the grid search for the random forest were: the minimum samples allowed in each split; the number of trees; and the maximum number of features used for each tree (Breiman, 2001). For the logistic regression the following parameters were tuned: the type of regression (characterised by different loss functions and regularisation schemes); the cost (or ‘C’); and the tolerance (epsilon). The grid search iterated through the following values for each meta parameter. For the type of regression: L2-regularised logistic regression (primal), L2- regularised L2-loss support vector classification (dual), L2- regularised L2-loss support vector classification (primal), L2- regularised L1-loss support vector classification (dual), L1- regularised logistic regression, and L2- regularised logistic regression (dual). Different cost values were: 1,2,3,4. The different Tolerance values were 0.0001,0.001,0.01,0.1,0.5.

10.3.2 Analysis 1: A list of the Non-Significant Predictor Variables in the Logistic Regression Model

- No. of children (continuous)
- Satisfaction with life (continuous)
- Conscientiousness (continuous)
- Extraversion (continuous)
- Agreeableness (continuous)
- Emotional Stability (continuous)
- Self-control (continuous)
- Shopping impulsivity (continuous)
- BAS-D (continuous)
- Shopping motivation question: ‘I would rather stick to a brand I usually buy than try something I am not very sure of’ (continuous)
- Shopping motivation question: ‘I enjoy taking chances in buying unfamiliar brands just to get some variety in my purchases’ (continuous)
- Shopping motivation question: ‘If I like a brand I rarely switch from it just to try something new’ (continuous)
- Shopping motivation question: ‘I would not mind paying more in order to get a high quality product’ (continuous)
- Shopping motivation question: ‘I only buy products I trust’ (continuous)
- Shopping motivation question: ‘Product Quality is extremely important to me’ (continuous)
- Shopping motivation question: ‘I am pleased about the way I look’ (continuous)
- Shopping motivation question: ‘When I feel good about my looks, I am happier and have a

better outlook on life' (continuous)

Shopping motivation question: 'Whenever I see a mirror, I "check myself out" to see how I look' (continuous)

Shopping motivation question: 'I am concerned about climate change (also known as global warming)' (continuous)

Gender (dummy variable)

Time preference: '£40 now or £70 in 3 months' (dummy variable)

Time preference: '£25 now or £70 in 3 months' (dummy variable)

Time preference: '£55 now or £70 in 3 months' (dummy variable)

Income: £100,000 to £149,999 (dummy variable)

Income: £150,000 or more (dummy variable)

Income: £25,000 to £34,999 (dummy variable)

Income: £35,000 to £49,999 (dummy variable)

Income: £75,000 to £99,999 (dummy variable)

Income: Less than £25,000 (dummy variable)

Income: 'Prefer not to say' (dummy variable)

Region: East Anglia (dummy variable)

Region: East Midlands (dummy variable)

Region: London (dummy variable)

Region: North East (dummy variable)

Region: North West (dummy variable)

Region: Northern Ireland (dummy variable)

Region: Scotland (dummy variable)

Qualification: A level (dummy variable)

Qualification: AS level (dummy variable)

Qualification: Advanced (dummy variable)

Qualification: BTEC (dummy variable)

Qualification: GCSE (dummy variable)

Qualification: Masters (dummy variable)

Qualification: Other (dummy variable)

Qualification: PhD (dummy variable)

Qualification: Postgrad (dummy variable)

Occupation: Casual worker (dummy variable)

Occupation: Full time carer of other household member (dummy variable)

Occupation: Higher managerial/professional/administrative (dummy variable)

Occupation: Homemaker (dummy variable)

Occupation: Other (dummy variable)

Occupation: Retired/Semi-retired (dummy variable)

Occupation: Semi-unskilled/manual work (dummy variable)

Occupation: Skilled manual worker (dummy variable)

Occupation: Student (dummy variable)

Occupation: Supervisory/clerical/junior/managerial/professional/administrative (dummy variable)

Occupation: Unemployed/long term sickness (dummy variable)

Marital: Divorced (dummy variable)

Marital: Other (dummy variable)

Marital: Separated (dummy variable)

Marital: Single (dummy variable)

People: Adults, family (dummy variable)

People: Adults, non-family (dummy variable)

People: Partner, children (dummy variable)

People: Solo (dummy variable)

Shopping frequency: Don't know (dummy variable)

Shopping frequency: Every 6 months or less (dummy variable)

Shopping frequency: Every day (dummy variable)

Shopping frequency: More than twice a week (dummy variable)

Shopping frequency: Once a month (dummy variable)

Shopping frequency: Once a week (dummy variable)

Shopping frequency: Once every 2-3 months (dummy variable)

Shopping frequency: Once every 4-6 months (dummy variable)

Shopping frequency: Hardly ever (dummy variable)

Card usage: Most of the time (dummy variable)

Card usage: Never (dummy variable)

Card usage: Sometimes (dummy variable)

10.3.3 Analysis 1: Structural Equation Model Pathway Analysis

Table 10.7: Table of results for the structural equation model predicting regular bag buying showing the regressions, covariances and all pathways.

Regressions	Standardised coef.	ci lower	ci upper	Z-value	p-value
Income ~Age (a1)	-0.022	-0.026	-0.018	-11.778	0.000
Impulsivity ~Age (a2)	-0.009	-0.010	-0.008	-16.533	0.000
Frugality ~Age (a3)	0.005	0.003	0.008	4.876	0.000
Bags ~Income (b1)	0.020	0.011	0.029	4.329	0.000
Bags ~Impulsivity (b2)	0.048	0.015	0.080	2.834	0.005
Bags ~Frugality (b3)	-0.090	-0.106	-0.074	-10.371	0.000
Covariances					
Impulsivity ~~Frugality	-0.021	-0.036	-0.006	-2.739	0.006
Income ~~Impulsivity	0.024	-0.010	0.058	1.598	0.110
Income ~~Frugality	-0.169	-0.239	-0.112	-5.577	0.000
Pathways					
Indirect 1: Bags~Age*Impulsivity	-0.000	-0.000	- 0.000	-3.657	0.000
Indirect 2: Bags~Age*Income	-0.000	-0.000	- 0.000	-2.862	0.004
Indirect 3: Bags~Age*Frugality	-0.000	-0.000	- 0.000	-4.073	0.000
Direct: Bags~Age	-0.011	-0.012	-0.010	-19.158	0.000
Total	-0.012	-0.014	-0.012	-22.137	0.000

10.4 Chapter 6: Understanding the Predictors of Plastic Bag Consumption in Big Data, Analysis 2

10.4.1 Analysis 2: Meta-Parameters

The meta parameter values considered in the grid search for the logistic regression were: the type of regression (characterised by different loss functions and regularisation schemes); the cost (or 'C'); and the tolerance (epsilon). The grid search iterated through the following values for each meta parameter. For the type of regression: L2-regularised logistic regression (primal), L2- regularised L2-loss support vector classification (dual), L2- regularised L2-loss support vector classification (primal), L2- regularised L1-loss support vector classification (dual), L1- regularised logistic regression, and L2- regularised logistic regression (dual). Different cost values were: 1,2,3,4. The different Tolerance values were 0.0001,0.001,0.01,0.1,0.5. This grid was finalised after ensuring an end value of any of the parameters had not been selected (i.e. the algorithm was not being limited by the parameter values provided).

10.4.2 Analysis 2: Full Results Table from the Logistic Regression Analysis

Table 10.8: Full results table from the logistic regression model predicting frequent bag buying.
 Table ordered by p -value from low to high.

	Independent Variables	Beta coefficient	p -value
1	Age	-0.600	0
2	Average Monthly Spend (£)	0.430	0
3	Gender: Male	0.210	0
4	Frugality	-0.210	0
5	Openness	0.140	0.001
6	Region: West Midlands	0.120	0.001
7	SM7	-0.130	0.001
8	Region: North West	0.120	0.002
9	Region: North East	0.110	0.002
10	Region: East Anglia	0.100	0.004
11	Shopping frequency: Every day	-0.090	0.007
12	Agreeableness	0.090	0.011
13	Region: London	0.080	0.019
14	SM6	0.090	0.020
15	Shopping frequency: Once in 2-3 months	-0.080	0.021
16	Income: Prefer not to say	-0.100	0.025
17	Shopping frequency: Once a month	-0.080	0.027
18	BAS-RR	0.100	0.030
19	Live with: Solo, children	0.080	0.032
20	Marital status: Divorced	0.080	0.036
21	Qualification: GCSE	0.090	0.037
22	Income: £150,000 or more	-0.070	0.038
23	SM9	0.070	0.046
24	Region: Yorkshire	0.070	0.047
25	Time preference: £40 now, not £70 in 3 months	0.100	0.049
26	Region: South East	0.070	0.054
27	Marital status: Widowed	0.060	0.066
28	Qualification: School	0.070	0.074
29	Marital status: Other	0.060	0.091
30	Shopping frequency: Twice a week	0.060	0.091
31	SM2	0.070	0.094
32	SM11	-0.070	0.096
33	Occupation: Skilled manual worker	0.060	0.110
34	Region: Wales	0.060	0.118
35	Live with: Partner, children	0.060	0.121
36	Region: East Midlands	0.060	0.125
37	SM3	0.060	0.127
38	Qualification: Postgraduate	0.060	0.143
39	Region: Northern Ireland‘	0.050	0.153
40	Shopping frequency: More than twice a week	0.050	0.167
41	Card usage: Never	0.070	0.169
42	Occupation: self-employed/own business	0.050	0.175
43	Marital status: Single	-0.060	0.183
44	SM4	0.050	0.192
45	Qualification: PhD	-0.040	0.205
46	SM5	0.050	0.208

	Independent Variables	Beta coefficient	<i>p</i> -value
47	Time preference: £55 now, not £70 in 3 months	0.050	0.243
48	Card usage: Hardly ever	0.040	0.267
49	Well-being	-0.040	0.270
50	Extraversion	0.040	0.307
51	Qualification: AS-level	-0.030	0.314
52	Qualification: A-level	0.040	0.316
53	Shopping Impulsivity	0.050	0.317
54	Region: Scotland	0.040	0.318
55	Region: South West	0.030	0.319
56	Income: £35,000 to £49,999	-0.040	0.351
57	Occupation: Retired/Semi-retired	-0.040	0.357
58	Occupation: Full-time carer of other household member	0.030	0.392
59	SM1	-0.040	0.413
60	Shopping frequency: Every 6 months or less	-0.030	0.432
61	Qualification: Advanced	-0.030	0.440
62	Occupation: Homemaker	0.030	0.474
63	Occupation: Casual worker	0.020	0.508
64	Income: £25,000 to £34,999	-0.030	0.527
65	Card usage: Most of the time	0.020	0.568
66	Live with: Adults (non-family)	0.020	0.571
67	Live with: Adults (family)	0.020	0.620
68	Qualification: BTEC	0.020	0.636
69	Shopping frequency: Once every 4-6 months	-0.020	0.644
70	Income: £50,000 to £74,999	-0.020	0.647
71	Card usage: Sometimes	-0.020	0.651
72	Qualification: Masters	0.020	0.659
73	Marital status: Separated	-0.010	0.672
74	Emotional Stability	-0.020	0.672
75	Time preference: £25 now, not £70 in 3 months	-0.020	0.686
76	Occupation: Unemployed/long-term sickness	-0.010	0.701
77	BAS-D	-0.020	0.709
78	Occupation: Other	0.010	0.715
79	Occupation: Semi/unskilled manual work	0.010	0.716
80	Income: £100,000 to £149,999	-0.010	0.721
81	Conscientiousness	0.010	0.726
82	SM10	0.010	0.731
83	Occupation: Higher managerial/professional/admin.	0.010	0.769
84	Occupation: Student	-0.010	0.782
85	Self-control	-0.010	0.795
86	Occupation: Supervisory/clerical/junior manager./admin	0.010	0.824
87	SM8	0.010	0.835
88	Qualification: Other	0.010	0.851
89	Income: £75,000 to £99,999	0.010	0.888
90	Shopping frequency: Once a week	0	0.902
91	Shopping frequency: Don't know	0.090	0.904
92	BAS-FS	0	0.933

Note: Significance levels and β coefficients are rounded to three decimal places. Shopping motivation (SM) questions: SM1= ‘I would rather stick to a brand I usually buy than try something I am not very sure of.’ SM2= ‘I enjoy taking chances in buying unfamiliar brands just to get some variety in my purchases.’ SM3= ‘If I like a brand I rarely switch from it just to try something new.’ SM4= ‘I would not mind paying more in order to get a high quality product.’ SM5= ‘I only buy products I trust.’ SM6= ‘Product Quality is extremely important to me.’ SM7= ‘I am pleased about the way I look.’ SM8= ‘When I feel good about my looks, I am happier and have a better outlook on life.’ SM9= ‘Whenever I see a mirror, I ”check myself out” to see how I look.’ SM10= ‘Environmental considerations affect the products that I purchase.’ SM11= ‘I am concerned about climate change (also known as global warming).’

10.5 Chapter 8: Investigating a Behavioural Measure of Anxiety in Transaction Data

10.5.1 Full List of Over the Counter Anti-Anxiety Medication used in the Study

Anxiety items:

- (5387299, “Rescue Remedy Night 10ml Dropper”, 2007), - sleep aid
- (5387302, “Rescue Remedy Night 20ml Spray”, 2007), - sleep aid
- (5629330, “Bach Rescue Remedy Pastilles Blkit 50g”, 2008), - natural stress relief sweets
- (6512704, “Kalms lavender one a day capsules 14s”, 2015), - herbal anxiety relief – “based on traditional use only”
- (5691664, “Kalms Blister 84 tabs”, 2009), - “relieves periods of worry, irritability, stresses and strains”
- (3869695, “Bach Rescue Remedy Pastilles 50g”, 2006), - natural stress relief sweets
- (6372961, “Rescue Plus food supplement lozenge 10s”, 2015), - b vitamins and flower essences
- (5966779, “Bach Rescue Remedy gummy stars 36s”, 2012), - sleep aids (2years+)
- (7518935, “Rescue Night Liquid Melts 28 capsules”, 2009), - sleep aid
- (1344560, “KALMS 200 tablets”, 1993), - sleep aid
- (2256827, “Schwabe Vitano 30’s”, 2008), - traditional medicine to relief stress and anxiety
- (2449226, “Tisserand de-stress aromatherapy roller”, 2008),
- (7709307, “BP St Johns Wort 250mg 30s”, 2010), - traditional medicine to a relieve depression
- (7707339, “Rescue Remedy Liquid Melts 28 caps”, 2012), - sleep aid
- (4115635, “Kalms day pocket pack tablet 42s”, 2011), - herbal anxiety relief
- (4153588, “Boots Stress Relief Tablets”, 2008), - rhodiola root extract
- (1454021, “Bach Rescue Remedy 20ML”, 1993), - ‘comfort and reassure’ dropper
- (1454056, “Bach Rescue Remedy 10ML”, 1993), - ‘comfort and reassure’ dropper

(4418786, “Pharmaton Capsules 30”, 2002), - vitamins and minerals for energy
(2821869, “Pharmaton Capsules 100”, 2005), - vitamins and minerals for energy
(7666543, “Kalms One a Night 21 tablets”, 2010), - sleep aid
(1850296, “Kalms night tablets 50s”, 2013), - sleep aid
(6217842, “Kalms day herbal tablets 100s”, 2014), - herbal anxiety relief
(5116279, “Kalms night tablets 21’s”, 2014), - sleep aid
(2319411, “Bach Rescue Spray”, 2002), - spray with flower essences
(1743430, “Rescue Cream 50ml”, 2016), - moisturiser with flower essences

10.5.2 A Full Table of Results from the Logistic Regression Model

Table 10.9: The full table of results from the logistic regression model predicting the purchasing of over the counter anxiety medication

	β coefficient (95% CI)
‘Age (years)’	0.01** (0.002, 0.03)
‘No. of children’	−0.15 (−0.60, 0.29)
‘£40 now or £75 in 2 months’	0.32 (−0.08, 0.73)
‘£25 now or £75 in 2 months’	−0.69*** (−1.18, −0.20)
‘£55 now or £75 in 2 months’	−0.01 (−0.34, 0.32)
Openness	0.09 (−0.03, 0.21)
Conscientiousness	0.08 (−0.05, 0.22)
Extraversion	−0.12** (−0.21, −0.03)
Agreeableness	0.0000 (−0.12, 0.12)
Emotional stability	−0.20*** (−0.30, −0.10)
Selfcontrol	0.07 (−0.18, 0.32)
Frugality	0.08 (−0.08, 0.24)
Shopping impulsivity	0.25*** (0.06, 0.45)
BAS-rr	−0.11 (−0.48, 0.26)
BAS-fs	−0.21 (−0.56, 0.14)
BAS-d	−0.05 (−0.30, 0.19)
‘I would rather stick to a brand...’	−0.07 (−0.17, 0.04)
‘I enjoy taking chances in buying...’	0.02 (−0.08, 0.11)
‘If I like a brand I rarely switch...’	0.05 (−0.05, 0.16)
‘I would not mind paying more...’	0.03 (−0.08, 0.15)
‘I only buy products I trust...’	−0.03 (−0.13, 0.07)
‘Product Quality is important...’	0.02 (−0.12, 0.16)
‘I am pleased about how I look...’	0.05 (−0.04, 0.14)
‘When I feel good about my looks...’	0.06 (−0.06, 0.17)
‘Whenever I see a mirror...’	−0.04 (−0.11, 0.04)
‘Environmental considerations...’	0.12** (0.02, 0.23)
‘I am concerned about climate...’	−0.10* (−0.20, 0.003)
Gender: Male	−0.81*** (−1.28, −0.34)
‘Qualification: GCSE’	0.18 (−0.21, 0.58)
‘Qualification: Masters’	0.04 (−0.45, 0.52)
‘Qualification: School’	0.20 (−0.44, 0.83)
‘Qualification: A-level’	−0.05 (−0.43, 0.33)
‘Qualification: BTEC’	0.14 (−0.42, 0.71)
‘Qualification: PhD’	−0.69 (−1.59, 0.21)
‘Qualification: Postgrad’	0.06 (−0.35, 0.48)
‘Qualification: Other’	−0.31 (−0.86, 0.23)
‘Qualification: AS-level’	0.26 (−0.91, 1.42)
‘Qualification: Advanced’	1.72 (−0.68, 4.13)
‘Region: West Midlands’	−0.03 (−0.62, 0.55)
‘Region: Northern Ireland’	−0.69** (−1.35, −0.04)
‘Region: Yorkshire’	−0.19 (−0.61, 0.22)
‘Region: Wales’	−0.16 (−0.69, 0.38)
‘Region: East Midlands’	−0.10 (−0.59, 0.39)
‘Region: South East’	−0.19 (−0.64, 0.26)
‘Region: South West’	−0.19 (−0.85, 0.48)
‘Region: East Anglia’	0.02 (−0.61, 0.64)
‘Region: North West’	−0.42 (−0.95, 0.11)
‘Region: North East’	−1.00** (−1.78, −0.21)
‘Region: Scotland’	−0.10 (−0.65, 0.44)
‘Region: London’	−0.44 (−1.03, 0.15)

	<i>β coefficient (95% CI)</i>
‘Occupation: Skilled manual worker’	−0.26 (−0.77, 0.26)
‘Occupation: Other’	0.16 (−0.70, 1.03)
‘Occupation: Supervisory/clerical/junior...’	0.12 (−0.23, 0.47)
‘Occupation: Higher manag./profes./admin.’	0.30 (−0.07, 0.66)
‘Occupation: Semi/unskilled manual work’	−0.53 (−1.23, 0.18)
‘Occupation: Casual worker’	0.05 (−1.34, 1.45)
‘Occupation: Retired/Semi-retired’	0.04 (−0.42, 0.49)
‘Occupation: Unemployed/long-term sick.’	0.12 (−0.81, 1.04)
‘Occupation: Full-time carer...’	0.03 (−1.22, 1.29)
‘Occupation: Homemaker’	−0.76 (−1.83, 0.32)
‘Occupation: Self-employed/own business’	0.39 (−0.83, 1.60)
‘Occupation: Student’	0.16 (−1.58, 1.90)
‘Income: £35,000 to 49,999’	−0.03 (−0.46, 0.41)
‘Income: £75,000 to 99,999’	0.53* (−0.10, 1.16)
‘Income: £25,000 to 34,999’	−0.08 (−0.48, 0.33)
‘Income: £150,000 or more’	0.01 (−0.99, 1.01)
‘Income: £50,000 to 74,999’	−0.16 (−0.63, 0.32)
‘Income: Prefer not to say’	−0.25 (−0.65, 0.15)
‘Income: £100,000 to 149,999’	0.12 (−0.69, 0.93)
‘Marital status: Single’	0.02 (−0.53, 0.58)
‘Marital status: Divorced’	−0.37 (−0.99, 0.25)
‘Marital status: Widowed’	−0.39 (−1.14, 0.36)
‘Marital status: Separated’	−0.44 (−1.37, 0.50)
‘Marital status: Other’	0.65 (−0.98, 2.28)
‘Living with: Adults, family’	0.04 (−0.36, 0.44)
‘Living with: Solo’	0.03 (−0.54, 0.60)
‘Living with: Partner, children’	0.23 (−0.58, 1.05)
‘Living with: Solo, children’	0.32 (−0.76, 1.40)
‘Living with: Adults, non-family’	−0.48 (−1.51, 0.55)
‘Shopping freq.: Once a week’	0.32** (0.001, 0.64)
‘Shopping freq.: Twice a week’	0.10 (−0.35, 0.55)
‘Shopping freq.: Once a month’	−0.31* (−0.64, 0.02)
‘Shopping freq.: Once every 2-3 months’	−0.43 (−0.95, 0.09)
‘Shopping freq.: More than twice a week’	1.07*** (0.39, 1.75)
‘Shopping freq.: Every day’	−0.50 (−1.99, 0.99)
‘Shopping freq.: Once every 4-6 months’	−2.17** (−4.11, −0.23)
‘Shopping freq.: Every 6 months or less’	−1.56 (−3.99, 0.86)
‘Shopping freq.: Don’t know’	−2.31** (−4.62, −0.004)
‘Card usage: Most of the time’	−0.52** (−0.93, −0.11)
‘Card usage: Sometimes’	−0.41 (−1.22, 0.40)
‘Card usage: Hardly ever’	−2.69 (−6.28, 0.90)
‘Card usage: Never’	−3.19 (−8.35, 1.97)
Well-being	−0.06 (−0.13, 0.01)
Observations	1,348
Log Likelihood	−837.06
Akaike Inf. Crit.	1,862.11
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01