Master Thesis

Grigor Dimitrov

11 October, 2016

Abstract

Write an abstract!

Contents

1 Introduction

This research explores the relationship between big five personality traits, risk attitudes and information search behavior in tourism context while controlling for trip characteristics and demographics. The paper offers unique contribution from methodological perspective as it is going to utilize behavioral datasets of the actual subjectsâ€™ online behavior with a combination of self-reported non-observable data. Up to date, there is a limited amount of literature exploring these constructs in relationship to one another and in a relationship with non-self-reported metrics of information sought.

1.1 Online behavioral data

The survey questionnaires are the main source of data for market research. Since the emergence of Web 2.0, the online survey made its way as it is a cheap and convenient form or data collection. With the increased penetration of the internet, there were more people online and this availability gave survey market research a huge reach. However, 20 years ago when the survey plunged our economic behavior wasnâ€™t focused that much online and the Internet was used mainly for information purpose (reference).

More than anytime nowadays everything happens online, our decision-making process is reflected online and online we can observe and capture information with ease. Consequently, there is an increasing availability of data. With passive data collection technologies, we can capture all of the online consumer behavior. Behavioral data which is more accurate in terms of describing actual behavior i.e. information gathering, research of alternatives and transactions which are the ultimate reflection of preference. Survey data is more powerful to collect information about respondentsâ€™ attitude, motivation, and opinion (reference).

The both types of data sources have their limitation and researchers are utilizing either one of them separately, however, up to date, there is a limited amount of research that combines both sources of data. By utilizing passive data collection panels this project has a rare opportunity to match an actual online behavior with a subject behind it and to send out questionnaires in order to access information that is not accessible via observational data.

The main motivation of the paper lays in the metodological contributions. Namely, utilization of observable online behavioral data collected via passive tracking technology regresses over subjects' unobservable characteristics collected via online questionnaire.

1.2 Tourism and The Web

Tourism is highly competitive and fragmented market. It has been disrupted by the Internet at large, the disruption process has started with the emergence of the Internet and continues today. We are witnessing dynamic segments of consumers emerged because of the technological advancements with constantly changing needs (Xiang, Magnini, and Fesenmaier 2015). Furthermore, there is an increasing availability of options and all sorts of services bringing the consumers and offers closer together manifesting the ongoing process of disintermediation. Xinag at al. (2015) notes that there is growing â€œbifurcationâ€ or a split among the traditional online travelers to users of traditional travel products and people how seeking deeper and authentic experiences. The authors point out that understanding how contemporary travelers use the internet is an important foundation for building successful communication strategies by the business stakeholders. The importance of the internet has been attributed to three main factors, the extensive amount of travel related information (ref), the development of social networks and travel related social networks where user can exchange travel experiences (ref) and the mobile computing, smartphones in particular (ref).

1.3 Personality traits

Big five inventory (BFI) is the most well-known and established factor structure to measure personality (Denissen, Geenen, van Aken, Gosling, & Potter, 2008). It has been validated on many occasions and widely utilized by researchers creating prerequisites to comparing finding across different studies. Another important factor for the popularity of BFI is the fact they are freely available to use along with their validated translations in many languages. BFI consists of Extraversion, Neuroticism, Conscientiousness, Agreeableness and Open to experience. These five factors are shown to be the main factors that drive the human behavior, appear in different cultures (Mowen 2000 are considering them universal), are relatively stable across the lifetime of subjects and have strong predictive validity. (reference).

Researchers using BFI often aim to measure big five personality dimensions using as less as possible items which lead to the development and validation of short scales of BFI (reference) shorter than the 44 item scale initially proposed by (reference).

Personality traits and tourism

According to (Leung & Law, 2010) the usage of personality traits in travel related literature appears to be low even though the academic agreed upon their value in Marketing domain (Baumgartner, 2002). In tourism domain, BFI has been examined with regards to, general travel behavior, the undertaken activities in the travel destination, with relationship of adventure tourism, pilgrim tourism (Jani, 2014).

Personality traits and tourism information search

There are two articles researching internet search behavior from the the perspective of BFI. Jani, (2011) proposes a model relating information needs and tourist information behavior from the perspective of BFI and travel personality. They define touristsâ€™ information needs as reasons for collecting information and information search behavior as breadth and depth of information sources consumers use for obtaining information.

Furthermore, Jani et al. (2014), research the relationship between BFI and internet search behavior. Using self-reported survey data, they observe that travel related information sought online varies with regards to the different personalities. Furthermore, the authors conclude that some of the factors from the BFI can improve the information search behavior predictability. Risk attitude Risk attitude and tourism Decisions related to travel involve a lot of risk and uncertainties. In an exploratory analysis, Roehl & Fesenmaier (1992) identify seven types of perceived risks related to pleasure travel. Namely, equipment risks, financial risks, physical risks, psychological risks, satisfaction risks, social risks and time risks. The team also refers to numerous studies suggesting the fact that perceptions toward risk are situation specific thus they should be studied with the appropriate measure for the specific context.

Lepp & Gibson (2003) look at the general perceptions or risk and uncertainty from the perspective of travel motivations for novelty against familiarity. Their findings suggest that familiarity seekers are more risk averse comparing to less risk averse novelty seekers. Risk attitude and tourism information search Quintal et al. (2010) examine the impact of risk and uncertainty avoidance on tourist information search. The team distinguishes risk from uncertainty and examine whether the two constructs have different impact on the information search. Their findings suggest that uncertainty avoidance has a positive significant relationship with the extent of information search while this is not valid for the risk avoidance.

1.4 Research question

> This research aims to answer the following research questions:
>
> What is the relationship between the different risk attitudes and the activity on travel realted domains? What is the relationship between the different uncertainty attitudes and the activity on travel realted domains?

The following measures are available for activity on travel realted domains:

- total number of travel related micromoments
- depth of search (total number of page views)
- breadth of search (total number of sources/websites used while searching)
- total time spent on websites associated with tourism related content

Doe to problems associated with time measurement on mobile there is an additinal measurement:

- total lenght of travel related micromoments

  Additional research:

  What is the relationship between the different personality traits and the activity on travel realted domains?

This research contributes with combining literature streams that have not been researched jointly i.e. BFI, decisions under risk and uncertainty and information search behavior in tourism domain. Furthermore, this research offers unique contributions in terms of data, i.e. combination self-reported non-observable data with behavioral (observable) data. The finding will be relevant to the current landscape of online products related to tourism because there an increasing availability of data from consumersâ€™ profiles which can be used to profile respondents in order to deliver personally customized tourism products. Personalization is a key element of the competition of the online tourism related services.

2 Literature review

2.1 Travel planning and information search

Tourism related information search and planning has been a widely researched topic. In a paper exploring longitudinal data for 12 years, Xiang, Magnini, and Fesenmaier (2015) identified as a key trend number one that the internet penetration among consumers using the Internet for travel planning already reached the level of saturation. However, in a research exploring the online behavior from a generation perspective, (Kim, Xiang, and Fesenmaier 2015) noted that there is no sufficient amount of research with regards to how different segments of consumers behave online and use the internet for travel planning. Internet search behavior in tourism context has been examined primarily from the perspective of demographic variables, motivation and prior knowledge about the destination (Jani, Jang, and Hwang 2014).

Xiang et al. (2015) turn special attention on travel planning process as a specific type of information search that is an important component of the decision making for tourism related decisions during which consumer are choosing their destinations and forming their expectations. There is a substantial amount of research that looks at travel planning from different perspectives some aim to identify the characteristics of the travelersâ€™ demographic characteristics (ref) other investigate the way travelers are conducting purchases and navigate in the information stream (ref) and recently social media and its influence on travel has been on focus of the researchers as well (ref).

From practitioner standpoint, the more consumers are being active online the more prerequisites this creates for the tourism stakeholders marketeers to reach them during their decision making process. Therefore, identifying the travel planning process i.e. the decision making journey is a critical step for the brands to intervene and influence the consumers in their direction, identifies Google travel survey (2014). Practitioners define that the decision making process related to the tourism as an array of micro moments which often are not even conscious to the consumers. (Reference to google travel survey) Micro moment is essentially a user session with particular goal of obtaining information or committing to a purchase. Again the authors stress out on the importance of mobile.

For the purpose of this research it is important to define the main components of the travel planning. In a meta-analysis, Jun, Vogt, and MacKay (2007) point out that the consensus among researchers that travel planning cannot be simplified to a single goal oriented rational action but it is rather viewed as a complex task involving multiple goals and decisions around the different goals and characteristics of the trip. The authors define a conceptual model for travel planning which has three main sequential interrelated components, pre-trip, during trip and post-trip. This research focuses on pre-trip phase. Pre-trip phase itself consists of information search and planning (decision making), furthermore, travel related purchases also occurs at the end of this phase as well as during the trip itself.

Jun, Vogt, and MacKay (2007) define travel plan as a complex decision involving an assessment of multiple alternatives organized around the travel goals in mind. The planning process includes setting goals and considering alternatives in order to achieve that goals including an evaluation of different alternativesâ€™ outcomes. Planning is dependent on the all information search behavior, utilization of the obtained information, purchase behavior and activities including past experience. Pan and Fesenmaier (2006) defines vacation planning over the internet as an interaction between the user and the â€œonline spaceâ€ related to destinations and tourism. The online space contains content provided by diverse sources and the technology that facilitates the communication. Userâ€™s â€œsituation, knowledge and skillsâ€ combined with the â€œonline spaceâ€ contribute to the effective search.

It is important to be noticed that trip planning is an important and enjoyable part of the vacation experience itself (Stewart and Vogt 1999) and at is likely to be high costs and high involvement decision (Bonn, Furr, and Susskind 1998). Furthermore, travel information search behavior explains travel purchase behavior (Woodside and MacDonald (1994)). Quintal, Lee, and Soutar (2010) review numerous aspect of from which information search has been researched including amount of search, number of sources, the search process, involvement, socio-demographic differences, culture etc.

Pan and Fesenmaier (2006) review the consumer vacation planning process from micro level perspective. The research is motivated by the fact that previous research has been mostly focused on exploring planning and information search on macro level i.e. motivation, need, determinants and outcomes. This research from micro level perspective focuses on a â€œsnapshotâ€ of travel planning where subjects make choices regarding a hypothetical holiday trip to San Diego. Using such setting it was possible to observe different chapters containing many episodes on how consumers adapt in their online search and come up with final decision.

In conclusion information search behavior is an important phase of the of the overall tourism behavior and more specifically travel planning. For tourism stakeholder perspective it is an crucial moment of where the consumer can be influenced with effective communication strategies and communication systems.

2.2 Risk and uncertainty

Risk attitudes are a central part of the economic theory. Classical economic theory of decision under risk states that the risk is related to the probability of the occurrence of specific outcome. For example according to expected utility a prospect with probability P to win x amount of money opposed to 1-P to win nothing, will be evaluated as follows: p.u(x)+(1-p)u(0) where u is the utility function of money. Risk attitudes are defined as follows, risk aversion is an attitude which is manifested by the preference of the sure outcome over a prospect with higher expected value that involves risk. Whereas, risk seeking attitude will occur when the prospect is preferred over the sure amount. Later economic theory evolves by distinguishing individual level probability weighing and utility weighting by taking into account different psychological variables. Kahneman and Tverski (1979) propose â€œProspect theoryâ€ to explain choices among risky prospects that are inconsistent with the standard economic theory. In these recent developments the risk attitudes evolve. In order to explain decision making under risk scholars explore choices involving different amount of risk (high risk and low risk) and associated with outcome involving different monetary values (again high and low) as well gains or loses.

In a paper focused on risk measurement in consumer research Mandrik and Bao (2005) summarize that the measurement of risk attitudes typically has been assessed in three ways. The first method involved â€œchoice dilemmasâ€ where subjects are presented with several scenarios and asked for their preference between two courses of action, this results in computing an overall score which is used to determine respondentsâ€™ risk attitude. The second method involves gambles. Subjects are asked to choose an amount in order to participate in a gamble. Finally, researchers use self-reported measures. These include creation of different scales that are measure risk and uncertainty in specific decision situations. The authors validate a novel self-reported scale which measure general risk attitude as valid psychometrical measure. The construct proposed by Mandrik & Bao has been utilized in this research as it provides shorter and simple manner of assessing risk attitudes. (argue that objective of the paper is to keep low length of the survey as low as possible)

2.3 Risk and uncertainty in tourism

One of the reasons researchers claim to cause the extensive information search is risk and uncertainty minimization. Stewart and Vogt (1999) attribute uncertainty as an implicit and universal characteristic of every planning process. Furthermore, the authors argue that in order to handle uncertainty the travelers prepare more than one plan for their trips. (Sweeney et al., 1999) point out that consumers who are more sensitive to risk and uncertainty engage in more extensive search in order to avoid them. (Sirakaya and Wood- side, 2005 claim that because of the intangible nature of tourism products the uncertainty in tourism is higher than comparing to other products or services. Based on this one of the hypothesis of this research has been formed, namely that the extend of information search is dependent on the attitudes toward risk and uncertainty.

An important remark related to risk and uncertainty is the difference between both constructs. The difference between them lays in the probabilities of their outcomes, while risks has been associated outcomes with known probabilities this is not the case with the uncertainty. Quintal, Lee, and Soutar (2010) are investigating the difference between risk and uncertainty on country level using Hofstedeâ€™s (1980) uncertainty avoidance index (UAI) and risk scale measurement on Touristâ€™ information search. The team claims that many other papers do not make the distinction and this is especially problematic when researchers are using country UAI scores to explain individual level behavior because individuals differ with regards to their attitudes of risk and uncertainty.

Quintal, Lee, and Soutar (2010) explain the relationship between uncertainty and risk in tourism and information search in the following way. In the early stages of their research consumers search for information extensively and the outcomes are more associated with uncertainty because the rate of occurrences of certain treating events is not known. In a later stage of decision making process, when travelers have already selected possible alternatives the risk attitude is more likely to have influence as consumer can assign relative probabilities i.e. alternatives are being compared to one another providing a reference point.

2.4 Big Five Factors (BFF)

Personality is a temperament or personâ€™s inherent qualities of mind and strategies according to which one behaves, dispositions and behavioral patterns that are stable across time and can be used to characterize oneâ€™s behavior. The trait perspective has been frequent utilized in consumer research because their ease of application as a self-reported measure and the measurement outputs can be easily applied in statistical analysis. (Jani, Jang, and Hwang 2014).

BFF has been proposed as a fundamental lexical hypothesis by Galton in 1884 (Golberg 1993), which is a language taxonomy of human temperament based on adjectives describing different personality traits. The theory was put into practice by Alport and Odbert (1936) and it has been grately developed ever since, leading into the construction of five broad factors. BFF are based on factor analysis where a large group of traits is shown to be correlated and grouped into five universal traits. BFF are openness to experience, conscientiousness, extraversion, agreeableness and neuroticism.

Openness to experience is related to the degree of curiosity, inventiveness, adoption of novelty on the one hand and consistency and cautiousness on the other. That is, persons with high openness tend to be open-minded, adventurous while low openness can describe individuals that are more pragmatic. Conscientiousness reflects on the tendency for one to be organized, non-spontaneous, organized and efficient. Extroversion is related to traits such as outgoing personality, sociability, talkativeness. Personalities exhibiting low extroversion on the other hand, can be perceived as less open and reserved. Agreeableness is described as being more compassionate and cooperative. Moreover, it measures whether a person can be trusted or not and if they are well-tempered. High agreeableness personalities can be seen as more naÃ¯ve, while low are seen as more dominative and competitive. Finally, neuroticism explores the emotional stability of individuals. That is, a high need for stability results in individuals who are clam and stable, while low need for stability can describe emotionally unstable individuals

2.5 Big Five and Touristsâ€™ internet search

Jani, Jang, and Hwang (2014) address the research question whether BFF can be used as predictor of internet search behavior in terms of sources of information and the extent information sought. Using self-reported measures of type of internet information search and channels used, the authors confirm that personality traits can be used as a predictor of information search behavior.

2.6 Hypotheses

Based on the literature review above I hypothetise the following relationship between the obseved behavior and the unobservables collected via survey data:

	Hypotheses			
H1	Risk attitude	*Decreases* the amount of	a.	Micromoments
			b.	Domains

Hypotheses				
			c.	Pageviews
			d.	Time
			e.	Lenght
H2	Uncertainty attitude	*Decreases* the amount of	a.	Micromoments
			b.	Domains
			c.	Pageviews
			d.	Time
			e.	Length

3 Methodology

3.1 Introduction

The data utilized for this research has been collected from panelists who participate in a large consumer panel in the Netherlands. The data has two main components, behavioral data and survey data. The methodological section of this paper explains the sampling, data collection and analysis procedures that have been conducted, in the following way:

Panel and sampling. Sampling based on behavioral data. A sample of respondents has been pulled out of the behavioral dataset based on their online behavior on popular tourism related websites in the Netherlands. Survey. An online survey has been administered among the sampled panelists to reveal the â€œunobservablesâ€ from the perspective of the available behavioral data. Namely, these include data points related to trip characteristics associated with respondentsâ€™ latest tourist related purchases as well as question regarding their personality. Data processing.The full behavioral data of the respondents who successfully completed the survey has been sampled out of the full behavioral panel dataset. Categorization. All of the unique websites of the sampled behavioral dataset have been classified into categories. In such a way it was possible to assess whether a certain website was travel related or not. Analysis. The aggregated information from the behavioral data has been regressed over the survey data to reveal the impact the unobservable personality attitudes and traits on the tourism related online behavior while controlling for the trip characteristics.

In the following sections I will first focus on explaining the behavioral data and the technology behind it. Afterwards, the 5 major steps of the methodology are explained in details. Then I proceed with the descriptive results of each of the datasets (survey and behavioral). Finally, I report the analysis results.

3.2 Behavioral data & Technology

The behavioral data, also referred to as observational data, reflects the online behavior of the consumers. It consists of records of the interaction made via consumersâ€™ digital devices and the Internet.

The behavioral data has been collected via a technology developed and provided for this research by Wakoopa. The company is a provider of a tracking technology. The technology is utilized primarily for market research purposes. Similarly to the market research consumer panels, where panelists enroll to participate in online surveys for incentives, Wakoopa provides its technology to market research consumer panel companies that are interested in tracking the online behavior of their panelists. After enrolling into the panel and giving their consent to be tracked, panelists install an application on their devices i.e. desktop, mobile and tablet. The tracking software collects every interaction of the panelistsâ€™ devices on the Internet which consists of path or the address the panelists are reaching and the duration of the visit. The software works in different manner over the different operation systems platforms and devices, but the final result consists of recording raw data containing events. Each event has the address a participant accessed, the duration of the interaction and the client requesting the information i.e. browser, app etc.

3.3 Steps

3.3.1 Panel and sampling

Using Wakoopaâ€™s panel it was possible to â€˜prescreenâ€™ relevant respondents for the purposes of the analysis, that could be invited to participate in the survey I have conducted. The respondents that I was looking for, should have been active on tourism related websites and should have conducted a purchase on such websites in the period of January 2015 to June 2016. Herein, I describe the process that was used to reach such subjects.

Upon starting of the project the panel used for the research had 6682 active panelists and 7103 active devices. The majority of the panelists were being active only on desktop. See appendix for reference. LINK

First, I started looking into tourism related websites in the Netherlands. Initially I used 300 domains to account for the majority of the tourism related internet traffic in the Netherlands according to the internet analytics company SimilarWeb. Afterwards, based on this data I exported the activity of the whole panel over those websites ranging from 01/2015 to 06/2016. The data of this activity was manually analysed looking for the end pages of the payments, also referred to as â€œconfirmationâ€ pages or â€œthank youâ€ pages. Confirmation pages are the pages where a customer has been redirected after conducting a purchase at a company website. In general, the analysis consists of looking up keywords within the travel domains URLs and marking down the common patterns with the aid of regular expressions (link to REGEX). Identified were the patterns of â€œconfirmationâ€ pages and data for the participants visited from the period of 01/03/2016 to 16/06/2016 was exported. Using this information, it was possible to identify a sample composed of 949 respondents which were to be invited into the online survey as described in the second step below. (See table in the appendix).

Given the estimation of the incidence rate provided by the panel supplier, 20%, and the initially desired sample of 500 respondents more respondents were needed. Therefore, an additional random sample of 123 respondents was selected based on whether they were active on the 300 initial domains but without evidence for their purchases from the data.

3.3.2 Survey

The selected 1039 panelists were invited to participate in an online survey, that aimed to reveal their attitudinal and personal characteristics as well as the trip characteristics of their last travel. The fieldwork was conducted during the last week of June 2016. Out of the 1039 invitations sent 872 started the survey, which resulted in 495 completed interviews. (Note to self, move this to a separate cleaning section). The data was further cleaned by accounting for speeders and flatliners. Speeders were panelists that have finished the survey within less than half of the average length of the interview. Flatliners refer to respondents who have answered all of the grid questions in a straight line. Moreover, a few respondents attempted to participate in the survey multiple times and thus, they were excluded from further analysis. Thus, the final dataset consists of 426 observations.

The survey consisted of several parts. First, a screening criteria was used. It accounted for the number of travel related purchases since March 2016. Respondents with no travel related purchases were not allowed to further proceed. Then, subjects who have not purchased neither flight nor accommodation were also excluded from the questionnaire. Business only travellers were not relevant for the analysis, thus they were also screened out, leaving only panelists who have gone on a leisure trip or both on a leisure and business trip. The second part was a demographics section, which consisted of question regarding age, gender and income. The next part of the questionnaire, was in regards to the trip characteristics, which was adapted from Roehl & Fesenmaier, (1992) and contained information about moment of booking, destination, planning horizon, information sources used i.e. Internet, advice from friend and relatives, tourist information office, travel agent etc., products purchased online, duration of the trip or number of nights spent away from home, whether the destination was visited before and how many times, number of travel companions, indication whether there was a child(ren) on the trip and whether subjects visited friends or relatives during their trip.

The next two sections aimed to reveal more about the subjects personality and risk and uncertainty attitude. Items assessing risk and uncertainty originate from Quintal et al. (2009). Respondents reported their risk attitude on three item scale which the authors adapt from Donthu and Gillilandâ€™s (1996). Uncertainty attitude has been assessed on four items scale which authors adapt from Yoo and Donthuâ€™s (2002) and the scale is based on Hofstedeâ€™s (1980) UA items. I have chosen these scales as they are the shortest reliable scales for self-assessing the risk and uncertainty attitudes. Using such scales I can make sure to not tire respondents and keep them engaged. Personality traits related to openness, consciousness, extraversion, agreeableness and neuroticism have been assessed using the short scale from the big five inventory proposed by (reference). In addition to the ten item scale an eleventh item was added as XXXX doesnâ€™t perform well in the ten item scale. The eleven items were derived from a validated Dutch translation of the BIG5 inventory from Denissen et al. (2008).

3.3.3 Data processing

After having collected and cleaned all the data, I further proceed with data processing and analysis. All of the data processing and the analyzation tasks of this paper were done in R Studio. All of the code, including libraries can be seen in the appendix.

The data processing includes merging down the two main behavioral streams of data that originate from desktop and mobile. Additional to that it is needed to classify and derive all the tourism related domains that will be further used for analysis. This is an important step as it will allow to analysis what the relationship between the online travel behavior and the travellerâ€™s attitudinal characteristics are. Categorization of the domains was done using machine learning classification algorithms provided by uClassify.com. The domains from the desktop and mobile behavioral data and all of the mobile apps are further classified using keywords.

Next step is, converting the raw data set onto micromoments. They are also further used in the analysis as I examine the relationship between oneâ€™s micromoment frequency and subjectâ€™s risk and uncertainty attitude. Micromoments are essentially user sessions where users were active within a given moment of time. For example, if a panelist access a certain domain and spend five minutes on it and then become inactive for more than 3 minutes, this will result in a factor variable grouping all of the observation within those five minutes of activity. If a travel related website has been visited during the micromoment, a dummy variable is assigned to this moment to indicate this. Furthermore, purchases of travel related products were included in the data based on the initial dataset used for pooling the sample out the panel for the survey. The final dataset includes aggregated information about panelistsâ€™ activity over three main levels: top-level i.e. total activity, mid-level activity before and after purchase, and low-level micromoment level.

This section includes the data processing tasks done on the behavioral data including descriptive statistics and variable derivation, classification and aggregation on the different levels intended for further analysis.

Starting point of the analysis of the behavioral data include procession the data and rendering it in a format suitable for running the analysis. In its initial form the data has been exported in a format containing the following variables: for the desktop data (sample data included in the appendix link) and for the mobile data (sample data included in the appendix link). The next step is adding â€œhostâ€ variable to both datasets.

The host variable contains domains and the subdomains that are going to be used to for the categorization whether the domains were travel related or not. The next query on the data included identifying all of the micromoments in the data. First, the data has been subsetted on panelist_id level, sorted by the the timestamp â€œused_atâ€ and assigned into list where each element is the full data for each individual respondent. The sum of â€œused_atâ€ variables and â€œactive_secondsâ€ , the duration in seconds respondent spend on the page, were compared to used_at variable of the next observation. If differences larger than five minutes were found all of the variables prior to this difference were grouped together under a common factor variable.

Desktop dataset:

[1] "used_at" "host" "panelist_id" "url"
[5] "active_seconds" "browser_name"

Mobile dataset:

[1] "panelist_id" "device_id" "scheme" "url" "domain"
[6] "app_id"

3.3.4 Categorization

The categorization procedure includes using uClassify.com machine learning algorithms and also using keywords. The full activity coming from desktop and mobile devices resulted in 194,534 unique domains. The categorization algorithm has been responsible for classifying all domains that have more than ten visits or 47,818 domains in total works as follows. A web-scraper designed for these projects accesses a the collection of domains and collects all of the information on the page, then removes the HTML elements along with the punctuation and renders down the information only to a part that is visible to the website visitors. Then, it passes this information to an application programmable interface (API) that returns the probability of this text being into sixteen different categories including travel.

The full list of categories includes:

[1] “Arts and Entertainment” “Autos”
[3] “Business Finance” “Celebrity”
[5] “College” “Cooking”
[7] “Dating and Romance” “Exercise”
[9] “Fashion and Beauty” “Games”
[11] “Health” “Home Improvement”
[13] “News” “Parents and Family”
[15] “Technology” “Travel”

Websites with content with the highest probability to be travel were assigned value of a dummy variable 1 or 0 otherwise.

Due to technical and time contains the API service wasnâ€™t able to classify all of the domains. Therefore, the full dataset of unique domains has also been scanned for keywords. Keywords include travel, tourism, accommodation, hotels, flight etc. (full list of the keywords can be found in the code in the appendix). If any of the keywords appear within the domain name, the respective domain has also been assigned to the list of travel domains. The same keyword approach has been used for classification of the application on mobile devices, where the apps have been classified using their names.

3.3.5 Analysis

The following section contains an explanation of the main techniques performed during the analysis along with their assumptions, followed by the results of the ordinary least squares diagnostic tests of the restricted model. Once the functional form of the restricted model has been selected I proceed with variable selection in order to come up with the final model. Finally, I ran the diagnostics of tests over the final model again.

3.3.5.1 Factor Analysis

Factor analysis is a widely used technique used for for explaining the variance in several variables by smaller set of latent variables. As in the current case it is often used to consolidate several survey variables onto their â€œunderlyingâ€ factors in order to reduce the dimensionality of the data. Factor analysis groups variables together, that is, using a large amount of variables one can potentially reduce them to certain factors representing the latent underlying factors representing them by accounting the similar patterns in the variables. The intuition behind the anaysis is as follows. The analysis groups together observed, correlated variables into smaller groups of unobserved (latent) variables (Yong and Pearce 2013).

In this case I use factor analysis to to reduce the seven survey items rearding the risk and uncertainty attitude down to two constructs namely risk and unertainty. Also to reduce the eleven item scale of BIG5 to 5 factors representing each of on the of the five personlity traits.

3.3.5.2 Regression Analysis

For testing the hypotheses of this paper, regression analysis will be utilized. The regression model or ordinary least squares (hereafter OLS) is the â€œcornerstone of econometricsâ€ (Verbeek, p.7, 2008). It aims at explaining a variable, y, in terms of another variable, x. In other words, using OLS researchers are able to find how will y vary as x changes, the ultimate goal being to infer the causal effect x has on y. Using such models allows to find relationships between various variables, present the effect the independent variables, xi have on the dependent variable, y in order to be able to make predictions.

The general linear regression models is represented as follows:

$$y = \quad {}_0 + X_1 \ {}_1 + X_2 \ {}_2 + \quad + X_k \ {}_k +$$

Where: y is the dependent variable

X_1 to X_k are the independent variables, which explain y

${}_0$ is the intercept, indicating the expected value of y when all the independent variables are equal to 0

${}_1$ to ${}_k$ are the coefficients which determine the effect x has on y

is the error term

3.3.5.2.1 Goodness of fit and model selection

The standard measures of fit include the R-squared and the adjusted R-squared, which measures the variance that is explained in the model for the independent variable by the dependent variables. The measure can be interpreted directly. For example if the R-squared is equal to 0.45, it means that the variables included in the model explain 45% of the variation of the independent variable, y. The higher the value, the higher its predictive power. However, it should be noted that adjusted R-squared penalizes for the additional number of parameters. Thus, applying additional variables to the model, I should test if they are jointly significant in order to assess whether they are relevant or not in the model.This is typically applied by using the F-test (Wald test). Using both the R-squared, the overall F-test and applying the F-test to certain variables I can compare best which model fits the data best.

3.3.5.2.2 Akaike information criterion (AIC)

Model selection has been done over Akaike information criterion (AIC) introduced by Akaike (1974). AIC is a metric traditionally used for model selection. It compares the goodness of fit for a number of explanatory variables and penalizes for each additional explanatory variable.

3.3.5.2.3 BLUE Assumptions

There are several assumptions that need to be met when applying OLS explained in the section below. Namely Gauss-Markov assumptions for full ideal conditions for OLS. The model needs to be best linear unbiased estimatorâ€ (â€œBLUEâ€) (Verbeek, 2008). It is crucial for the assumptions to be met as to compute unbiased and consistent estimates that explain the variation in the dependent variable. Now, I will go through each assumption: Linear in parameters This implies that the model should have linear parameters, b, however, there can be nonlinearities in the variables, x. This assumption is met as my specified model does not include non-linearities in the parameters.

3.3.5.2.4 Normality

The error termâ€™s should follow a normal distribution. In large datasets, however, even if the error term does not follow a normal distribution the regression estimators are â€˜asymptotically normally distributedâ€™, meaning that following non-normal distribution is not crucial as the estimates will still be consistent and unbiased. The

Shapiro-Wilk test can is adopted here and results presented below. The test works under null hypothesis: â€œthe sample comes from normally distributed populationâ€ Shapiro and Wilk (1965)

3.3.5.2.5 Random sample

The data collection should be done randomly, meaning that the each subject should have the same probability of being selected. In this research, both in the behavioral and survey data collection parts, I can say that subjects were randomly selected for further analysis.

3.3.5.2.6 Multicollinearity

Multicollinearity implies that there is no perfect linear relationship between the independent (explanatory) variables as this can lead to â€˜unreliable regression estimatedâ€™ (Verbeek, p. 42, 2008). For example, adding both male and female in the analysis would lead to perfect collinearity (as male + female =1) and the estimations would not work. In this example, removing one of the variables would solve the problem, however there can be other variables that are highly correlated. Having multicollinearity would not lead to biased estimates, but to inaccurate estimates. In such a case, excluding variables from the model should be considered. There are no tests that specifically look for multicollinearity, however there are certain indications. For instance, having two variables that are jointly significant (have big F-statistics), but independently are not significant can be a sign of multicollinearity

3.3.5.2.7 Homoscedasticity

Homoscedasticity implies that the variance of the error term should be the same for all values of the independent variables. If this does not hold, there is problem with heteroscedasticity meaning that the estimates of the regression are inconsistent due to inaccuracy of their standard errors, meaning that the t-statistics and thus the significance level of the estimates is not valid anymore. To test for homoscedasticity, I perform the Breusch-Pagan test, which hypothesize that there is constant variance of the error terms. Endogeneity The last assumption is crucial to be met as otherwise the regression estimates are biased and inconsistent. Endogeneity implies that there is correlation between an independent variable and the error term. There are several reasons why this assumption does not fold: 1 The model is misspecified. That is, nonlinearities are missing from the model or interaction effects are not accounted for. To account for that I perform the Ramsey-Reset test. The tests adds fitted values on power and re-estimates the model. The intuition behind it is that if non linear combination of independent variables can explain the dependent variable there are evidence the model is misspecified. The Ramsey-Reset test work under null hypothesis that the model has no important omitted non-linearities (Ramsey 1974) 2 Endogeneity, meaning that we are either missing important variables that explain the variance in the independent variables or we have reverse causality, that means that there can be a loop of causality between the independent and dependent variable.

3.3.5.2.8 Stepwise regression

The idea of stepwise regression has been introduced by Hastie and Pregibon (1992) and further improved by Ripley (2002). It is an iterative function ran over a restricted model and a set of candidate models. Each candidate model consists of a different set of explanatory variables. The function computes iteratively Akaike information criterion (AIC) values for the models comparing them to the best performing models from the previous iteration and based on the performance chooses whether to continue the loop with the new model or remain with the old one. The final output is the best performing model.

3.3.6 Model selection

3.3.6.1 Restricted Model

There are two distinct sources of data resulting in three data sets to be investigated.

1. Dataset from desktop
2. Dataset from mobile
3. Combined dataset from desktop and mobile

Furthermore, there are five dependent variables that are intended to be examinated during this research. Namely,

1. Number of travel micro-moments

2. Number of unique travel domains visited
3. Number of travel pageviews
4. Total time in seconds spent on travel domains
5. Total length in seconds of travel micro-moments

The difference between the last two is as follows. Total time in seconds has been measured once a panelist arrives at a certain website that has been classified as travel, whereas length is the total length (i.e. the difference between the start and end) of micro-moments amongst which a participant visited a travel related domain.

The restricted models of each one of the three datasets take the following form:

Table 3.1: Restricted model

Model	Dependent variables	Independent variables	Control *	Control2 **
# 1	Number of travel micro-moments	Risk + Uncertainty + Interaction	Total micro-moments	Total purchases + Device + Days active in the panel
# 2	Number of unique travel domains visited		Total domains	
# 3	Number of travel pageviews		Total pageviews	
# 4	Total time in seconds spent on travel domains		Total time	
# 5	Total length in seconds of travel micro-moments		Total micro-moments length	

* As there was only one purchase detected on mobile devices this term wasnâ€™t present in mobile model

** The variable device is present only in the combined model. It has value 1 is a respondent has been active only on desktop 2. Only on mobile and 3 if the respondent has been active on mobile and desktop. Obviously, in the models containing only desktop and only mobile data the variable has no variance therefore it is excluded. Since the respondents are selected only based on their desktop behavior there are no respondents who has been active only on mobile. Therefore the device variable in the combined dataset has only values 1 and 3.

Running Ramsey-Reset test per each on of the basic models showed evidence that the functional form of the models is not well specified, thus the models were rejected. Consequently, all of the dependent variables along with their corresponding controls were transformed into logs accounting for additional five models per each dataset. The results from consequent running of Ramsey-Reset test still were not satisfying, as the test still showed significant results indicating that the functional form of the model is not well specified. Consequently, the corresponding control variables accounting for total activity and purchases were transformed into binary variables with values 1 indicating a participant belongs to a highly active groups of participants (having values above the mean of the sample) and 0 vise versa. The latter transformation accounted for investigating ten more models per each dataset reaching the total number of thirty models to be reviewed. The functional form of the model has been selected based on the dataset containing desktop data due to the following reasons: it is the most complete in terms of number of observation, the participants were selected based on their desktop behavior.

Based on Ramsey-Reset test, I select the following functional form of the model for Desktop, Mobile and Combined datasets. Shapiro-Wilk test for normality also performs best for this functional form, yet the sample is large enough so we can relax the normality assumption.

Model	Dependent variables	Independent variables	Control *	Control2 ** ***
# 1	log(#) of travel micro-moments	Risk + Uncertainty + Interaction	(D) Total micro-moments	(D) Total purchases + Device + Days active in the panel

Model	Dependent variables	Independent variables	Control [*]	Control2 [**] [***]
# 2	log(#) of unique travel domains visited		(D) Total domains	
# 3	log(#) of travel pageviews		(D) Total pageviews	
# 4	log(#) time in seconds spent on travel domains		(D) Total time	
# 5	log(#) length in seconds of travel micro-moments		(D) Total micro-moments length	

[*] Note: as there was only one purchase detected on mobile devices this term wasnâ€™t present in mobile model
[**] Note: The variable device is present only in the combined model. It has value 1 is a respondent has been active only on desktop 2. Only on mobile and 3 if the respondent has been active on mobile and desktop. Obviously, in the models containing only desktop and only mobile data the variable has no variance therefore it is excluded. Since the respondents are selected only based on their desktop behavior there are no respondents who has been active only on mobile. Therefore the device variable in the combined dataset has only values 1 and 3. [***] Dummy variables take value 1 is the respondent is activity is above the average activity of the sample and 0 otherwise.

A table with the results from the restricted model along with the model performance metrics can be found in the appendix:

```
##
## ==================================================================================
##                                                      Dependent variable:
##                          ----------------------------------------------------
##                          log(MM)  log(Domains) log(PV)  log(Time) log(Length)
```

```
##                              (1)        (2)        (3)       (4)        (5)
## ------------------------------------------------------------------------------------------
## Risk:seek                -0.16     -0.25**    -0.42*** -0.47***    -0.28**
##                          (0.12)     (0.12)     (0.13)   (0.13)     (0.12)
##
## Uncertainty:seek           -0.15     -0.26**    -0.31*** -0.32***    -0.21*
##                          (0.11)     (0.11)     (0.12)   (0.12)     (0.11)
##
## Days                    0.01***    0.01***   0.004*** 0.004***    0.01***
##                         (0.001)    (0.001)    (0.002)  (0.002)    (0.001)
##
## D MM                               1.17***
##                                     (0.10)
##
## D Domains                                     1.08***
##                                                (0.10)
##
## D PV                                                   0.88***
##                                                         (0.11)
##
## D Time                                                            0.86***
##                                                                    (0.11)
##
## D Length                                                          1.34***
##                                                                    (0.10)
##
## D Purchase               0.50***    0.62***    0.86***  0.87***    0.48***
##                          (0.11)     (0.11)     (0.11)   (0.11)     (0.10)
##
## Risk x Uncertainty          0.27       0.32       0.66**   0.83**      0.48
##                          (0.30)     (0.31)     (0.32)   (0.32)     (0.29)
##
## Constant                3.12***    3.59***    5.66***  8.94***   10.99***
##                          (0.27)     (0.27)     (0.29)   (0.29)     (0.26)
```

```
##
## ----------------------------------------------------------------------------------
## Observations                    426       426       426      426       426
## R2                            0.36      0.35      0.30     0.31      0.41
## Adjusted R2                    0.35      0.34      0.29     0.30      0.41
## Residual Std. Error (df = 419)   0.97      0.98      1.02     1.03      0.93
## F Statistic (df = 6; 419)     39.89***  37.21***  29.69*** 31.07***  49.37***
## ==================================================================================
## Note:                                         *p<0.1; **p<0.05; ***p<0.01
```

Table: Desktopr data, restricted model

```
##
## ==================================================================================
## Residual_Standard_Err F.Stat NumDF FDenDF R.Sq Adj.R.Sq Shapiro_Wilk_Stat Shapiro_Wil
## ----------------------------------------------------------------------------------
## 0.97             39.89   6   419  0.36  0.35      1.00           0.94       2.03     0
## 0.98             37.21   6   419  0.35  0.34      1.00           0.24       1.20     0
## 1.02             29.69   6   419  0.30  0.29      0.99           0.16       0.79     0
## 1.03             31.07   6   419  0.31  0.30      0.98          0.0001       1.31
## 0.93             49.37   6   419  0.41  0.41      0.99           0.01       1.97     0
## ----------------------------------------------------------------------------------
```

Table: Desktopr data, restricted model tests

```
##
## % Error: 'style' must be either 'latex' (default), 'html' or 'text.'
```

Table: Mobile data, restricted model

```
##
## ==================================================================================
## Residual_Standard_Err F.Stat NumDF FDenDF R.Sq Adj.R.Sq Shapiro_Wilk_Stat Shapiro_Wil
## ----------------------------------------------------------------------------------
## 75.41             1.89   5    95  0.09  0.04      0.46            0          0.41     0.
```

```
## 211.98          42.72   5    95   0.69   0.68       0.65           0        205.85
## 59.27           10.91   5    95   0.36   0.33       0.56           0        25.23
## 106,753.30        5.67   5    95   0.23   0.19       0.38           0        15.69    (
## 157,353.90        3.87   5    95   0.17   0.13       0.40           0         3.63
## ----------------------------------------------------------------------------------------
```

Table: Mobile data, restricted model tests

```
##
## ========================================================================================
##                                          Dependent variable:
##                        ----------------------------------------------------
##                        log(MM)  log(Domains) log(PV)  log(Time) log(Length)
##                              (1)        (2)        (3)       (4)        (5)
## ----------------------------------------------------------------------------------------
## Risk:seek                  -0.20     -0.20    -0.35*** -0.38***    -0.19
##                           (0.13)    (0.13)    (0.13)   (0.14)     (0.15)
##
## Uncertainty:seek            -0.13     -0.15     -0.23*   -0.19     -0.17
##                           (0.12)    (0.12)    (0.13)   (0.13)     (0.14)
##
## Days                    0.01***   0.01***   0.01***  0.01***   0.01***
##                         (0.001)   (0.001)   (0.002)  (0.002)   (0.002)
##
## D MM                             0.91***
##                                   (0.11)
##
## D Domains                                  1.01***
##                                             (0.10)
##
## D PV                                                0.82***
##                                                      (0.11)
##
## D Time                                                         0.77***
##                                                                 (0.12)
```

```
##
## D Length                                                      0.98***
##                                                               (0.12)
##
## D Purchase                0.59***   0.62***   0.90***  0.85***   0.51***
##                            (0.11)    (0.11)    (0.12)   (0.12)    (0.13)
##
## Risk x Uncertainty            0.25      0.27     0.58*    0.58      0.48
##                            (0.32)    (0.31)    (0.34)   (0.35)    (0.37)
##
## Constant                  2.67***   3.16***   5.08***  8.10***  10.01***
##                            (0.29)    (0.28)    (0.30)   (0.31)    (0.32)
##
## ------------------------------------------------------------------------
## Observations                   429       429       429      429       429
## R2                            0.30      0.34      0.30     0.29      0.29
## Adjusted R2                    0.29      0.33      0.29     0.28      0.28
## Residual Std. Error (df = 422)  1.03      1.00      1.07     1.13      1.16
## F Statistic (df = 6; 422)    30.57***  36.53***  29.52*** 29.22***  29.27***
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

Table: Combined data, restricted model

```
##
## ========================================================================
## Residual_Standard_Err F.Stat NumDF FDenDF R.Sq Adj.R.Sq Shapiro_Wilk_Stat Shapiro_Wil
## ------------------------------------------------------------------------
## 0.87              70.65   6    422   0.50   0.49       1.00           0.36       10.42    0
## 0.81              90.89   6    422   0.56   0.56       0.98           0.0001      5.80
## 0.91              68.73   6    422   0.49   0.49       1.00           0.28        2.72    0
## 0.97              62.99   6    422   0.47   0.46       0.97           0.0000     12.86
## 1.06              49.46   6    422   0.41   0.40       0.75             0        1.30    0
## 267.80            12.88   6    422   0.15   0.14       0.73             0        2.48    0
## ------------------------------------------------------------------------
```

Table: Combined data, restricted model tests

3.3.6.2 Final Model

Based on the selected functional form of the model(s) I proceed with stepwise selection in order to select the final model across desktop, mobile and combined dataset. First, the models are presented then results of the OLS performance tests are shown and discussed.

Results: https://www.dropbox.com/s/x6y4xl3sh4elr3u/Screenshot%202016-09-15%2018.58.17.png?dl=0 https://www.dropbox.com/s/fas9fw1y5jgo1pz/Screenshot%202016-09-15%2018.58.34.png?dl=0 https://www.dropbox.com/s/axaanyn7mfpctbt/Screenshot%202016-09-15%2018.58.45.png?dl=0

Tests: https://www.dropbox.com/s/dsl9e1zhton9w02/Screenshot%202016-09-15%2018.57.21.png?dl=0

4 Restults

4.1 Survey Data

The original data consisted of 495 observations. Before proceeding with analysis, the data was cleaned based on several considerations. Namely, I looked at whether 1) respondents were speeding throughout the survey, 2) subjects were flatliners, e.g. if they have answered the grid question with the same answer, and 3) if they had low activity less than 1000 observation in the behavioral data. 4) if the attempted to participate the survey more than once. The final dataset thus, consists of 426 observations.

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Oct 11, 2016 - 5:49:53 PM

Looking at the demographic characteristics, we see that there is equal split between males and females. Seeing the distribution among age, the population above 50 years old is overrepresented, accounting for 59% of the sample. In terms of income, the distribution is normal, most of the subject indicated yearly income between 33.000 â€“ 66.000 euro.

Table 4.1:

	raua
Q16_1	I would rather be ewline safe than sorry.
Q16_2	I want to be sure ewline before I purchase anything.
Q16_3	I avoid risky things.
Q16_4	It is important to ewline have instructions spelled out in detail so I always know what I am
Q16_5	It is important to ewline closely follow instructions and procedures.
Q16_6	Rules and regulations ewline are important because they tell me what is expected
Q16_7	Standardised work ewline procedures are helpful.

All subjects were actively involved in the trip planning as a main decision maker or contributed to the decision, subject who didnâ€™t participated in the decision making were screened out. Business trip only travellers were screened out as well, those left in the sample are either leisure or leisure and travelers. Finally, we see that the majority of respondents indicated they have booked both flight and accommodation.

The next part of the survey, explored the trip characteristics with regards to their general planning, online behavior and destination. These data points were used as control variables for the analysis. On average it took respondents 9 weeks to plan the whole trip, for about 40% it took between 1 and 4 weeks to plan everything, while 15% indicated they needed more than 20 weeks of planning.

The majority (80%) visited countries within Europe, and 57% have not visited the country before. On average, people who have visited the country before went there for the 4th time. Typically together were travelling 2 people, the largest travel group consisted of 14 people and they have spent approximately 9 nights away from home. About 34% of the respondents indicated they have spent more than 10 days. Only 20% of the subjects have stayed with relatives.

The majority of subjects researched for the trip using the Internet, but many have also asked for advice from friends or relatives or approached a travel agent. Moreover, the internet is widely used for booking accommodation and not so much for travelling or booking â€˜entertainmentâ€™ online.

4.1.1 Risk and uncertainty

For understanding risk and uncertainty attitudes the seven item scale was analyzed using factor analysis. According to the analysis results we can not accept the hypothesis that two factors are sufficient to explain the seven item scale that has been used. However exploring further the resuts show that the factor loadings of the different variables have the highest loadings on the assumed two underlying factors i.e. risk and uncertainty. Furthermore, running factor analysis when assuming 3 factors resuts in separating item #4 as a distinct factor with low factor loading. According to (Yong 2013) choosing too few factors may lead to leaving out importan valance out of tha anlysis. However, I accept the high factor loadings onto the two assumed factors as a sufficient reason to continue the analysis with two factors. Namely, items 1 to 3 accounting for underlying risk attitude and items 4 to 7 accounting for attitude towards uncertainty.

```
##
## Call:
## factanal(x = RAUAdata, factors = 2, rotation = "varimax")
##
## Uniquenesses:
## Q16_1 Q16_2 Q16_3 Q16_4 Q16_5 Q16_6 Q16_7
##  0.35  0.20  0.64  0.63  0.33  0.25  0.60
##
## Loadings:
##       Factor1 Factor2
## Q16_1         0.71
## Q16_2         0.88
## Q16_3         0.53
## Q16_4 0.49
## Q16_5 0.78
## Q16_6 0.84
## Q16_7 0.55
##
```

```
##                 Factor1 Factor2
## SS loadings       2.11    1.90
## Proportion Var    0.30    0.27
## Cumulative Var    0.30    0.57
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 29.15 on 8 degrees of freedom.
## The p-value is 0.000299
```

4.1.2 BIG5

Running exploratory factor analysis on BIG5 items resuled in confirming the initially assumed personality traits. Furthermore, based on the results we can accept the hypothesis that 5 factors are sufficient to explain the variance of the 11 item scale.

```
##
## Call:
## factanal(x = BIG5data, factors = 5, rotation = "varimax")
##
## Uniquenesses:
##  Q15_1  Q15_2  Q15_3  Q15_4  Q15_5  Q15_6  Q15_7  Q15_8  Q15_9 Q15_10
##   0.70   0.49   0.68   0.52   0.60   0.61   0.70   0.21   0.63   0.80
## Q15_11
##   0.48
##
## Loadings:
##        Factor1 Factor2 Factor3 Factor4 Factor5
## Q15_1                          -0.52
## Q15_2                           0.69
## Q15_3                                   0.49
## Q15_4                                  -0.63
```

```
## Q15_5                  0.61
## Q15_6                  0.60
## Q15_7  -0.40
## Q15_8   0.89
## Q15_9           0.60
## Q15_10
## Q15_11          0.69
##
##                Factor1 Factor2 Factor3 Factor4 Factor5
## SS loadings       0.99    0.97    0.97    0.91    0.74
## Proportion Var    0.09    0.09    0.09    0.08    0.07
## Cumulative Var    0.09    0.18    0.27    0.35    0.42
##
## Test of the hypothesis that 5 factors are sufficient.
## The chi square statistic is 4.15 on 10 degrees of freedom.
## The p-value is 0.94
```

Next, with the aid of factor analysis respondents were grouped into 2x2 groups: 1) risk & averse seeking and 2) uncertainty averse and seeking. It is interesting to observe some of the descriptive statistics which can already provide some insights on what the relationships between peopleâ€™s risk and uncertainty attitude and their travel behavior based on the survey data. First of all, the majority of subjects are risk averse and uncertainty averse. Subject being risk seeking account for 30% of the population, while the uncertainty seeking are 40%.

Uncertainty attitude

averse seeking Risk attitude averse 133 161 seeking 119 13 Note: The correlation between the means of RA and UA scales show correlation of 0.58

Interestingly enough, given the data, it is observed that risk averse and uncertainty avoidant subjects actually prefer to go to a destination they haven't visited before, while risk and uncertainty seeking respondents prefer a bit better to go to a known destination (see Table 4). Table 4, RA & UA attitude repeated visit

Prior visit of destination RA - averse RA-seeking Yes 38% 44% No 62% 56%

UA - averse UA-seeking Yes 38% 42% No 62% 58%

Further looking at the data, however, I find that on average risk averse people tend to spend more time planning the trip than risk reeking people (10 days vs 7 days). This effect is significant at 5% significance level. However, uncertainty averse and seeking people exhibit similar behavior (9 versus 10 days), the mean difference is not significant at 5% significance level, and thus there is no difference in the planning horizon for uncertainty averse and seeking travelers.

4.2 Main results

Results:

```
##
## ================================================================================
##                                                Dependent variable:
##                       ---------------------------------------------------
##                       log(MM) log(Domains) log(PV)  log(Time) log(Length)
##                           (1)      (2)        (3)      (4)        (5)
## --------------------------------------------------------------------------------
## Risk:seek                -0.11    -0.23*    -0.37*** -0.46***   -0.26**
##                         (0.12)    (0.12)     (0.13)   (0.13)     (0.12)
##
## Uncertainty:seek           -0.13   -0.25**   -0.34*** -0.37***    -0.17
##                         (0.11)    (0.11)     (0.12)   (0.12)     (0.11)
##
## Risk/Uncertainty:seek     0.01***  0.01***   0.01***  0.01***  0.005***
##                        (0.001)   (0.001)    (0.002)  (0.001)    (0.001)
##
## Days active              0.53***  0.63***    0.84***  0.88***   0.50***
##                         (0.11)    (0.11)     (0.11)   (0.11)     (0.10)
##
## D Purchase                    1.13***
##                               (0.10)
##
```

```
## D Micromoments                         1.03***
##                                        (0.10)
##
## D Domains              -0.20**  -0.28***  -0.33***  -0.26**  -0.26***
##                         (0.10)    (0.10)    (0.10)   (0.10)    (0.09)
##
## D PageViews                                         0.83***
##                                                      (0.11)
##
## D Time                                                        0.80***
##                                                               (0.11)
##
## D MM Length                             0.21*      0.21*     0.29**
##                                        (0.11)     (0.12)    (0.11)
##
## S1. 1 purchase                           0.15       0.20     0.28**
##                                        (0.12)     (0.13)    (0.12)
##
## S1. 2 purchase                                                 1.36***
##                                                                 (0.10)
##
## S1. 3+ purchase              1.01      2.58**     -0.52
##                             (1.02)     (1.02)     (1.08)
##
## S2. Primary purchase: Flight  1.87*     2.14**     2.01*
##                             (1.00)     (1.01)     (1.06)
##
## S3. decision/partly                -0.51     -0.30    0.08      0.33*
##                                   (0.35)    (0.36)  (0.36)     (0.18)
##
## D1. Age                             0.65      0.65    0.57     0.38**
##                                   (0.71)    (0.74)  (0.73)     (0.18)
##
## D1. Age Sq                         -0.01     -0.82  -1.13**     0.25
```

```
##                                  (0.50)    (0.52)   (0.52)    (0.25)
##
## Q1. Dest: Asia (B:Eur)            -0.64     -0.65    -0.36     -0.63
##                                  (0.57)    (0.60)   (0.60)    (0.42)
##
## Q1. Dest: NA (B:Eur)              -0.05      0.12     0.35     -0.14
##                                  (0.35)    (0.36)   (0.36)    (0.34)
##
## Q1. Dest: SA (B:Eur)        -0.28**   -0.38**    -0.37**   -0.31**
##                             (0.13)     (0.15)     (0.16)   (0.16)
##
## Q1. Dest: AUS (B:Eur)                               0.01
##                                                   (0.004)
##
## Q1. Dest: Africa (B:Eur)    -0.21**  -0.24**   -0.29*** -0.37***   -0.17*
##                            (0.10)    (0.10)    (0.10)   (0.10)    (0.09)
##
## Q4. Plan horizon (Weeks)     0.27*    0.36**    0.37**            0.38**
##                             (0.16)    (0.17)    (0.17)            (0.16)
##
## Q5.2. Advice of friends/rel  0.18*    0.20**     0.14   0.20**    0.19**
##                            (0.09)    (0.10)    (0.10)   (0.10)    (0.09)
##
## Q5.3. Tourist info office   -0.43*                                -0.41*
##                             (0.23)                                (0.22)
##
## Q5.4. Travel mag                                 0.43               0.39
##                                                (0.28)             (0.25)
##
## Q5.5. Travel agent            -0.16     -0.17*
##                               (0.10)     (0.10)
##
## Q6.1. Online: Transport                             -0.32     -0.33
##                                                    (0.22)   (0.22)
```

```
##
## Q6.2. Online: Accomm.        -0.27*     -0.25*    -0.34**    -0.23
##                              (0.15)     (0.15)     (0.15)   (0.15)
##
## Q6.3. Online: Entert.                                       -0.20**
##                                                             (0.10)
##
## Q7. Length: 3+ nights                              -0.40     -0.36
##                                                    (0.25)   (0.25)
##
## Q8. Visited before          0.09      0.10      0.39   0.76**     0.30
##                           (0.30)    (0.31)    (0.32)  (0.32)    (0.29)
##
## Q10. Numer of people                   1.16***     0.79*     0.54
##                                         (0.41)     (0.43)   (0.43)
##
## Q11.1. Children (N)
##
##
## Q11.2. Visited Fr/Rel (N)               -0.31      -0.46     -0.37
##                                         (0.73)     (0.76)   (0.76)
##
## Q11.3. Stayed at Fr/Rel (N)              0.68       1.09     0.55
##                                         (0.90)     (0.93)   (0.93)
##
## Q11.4. Stayed at Hotel (N)               0.32      1.12*    1.74***
##                                         (0.58)     (0.61)   (0.61)
##
## Q11.5. Group Trip (N)      3.91***  3.89***   5.36***  9.29***  10.69***
##                           (0.43)    (0.39)    (0.49)  (0.35)    (0.46)
##
## ===============================================================================
## ===============================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

```
##
## ===========================================================================================
##                                                          Dependent variable:
##                                 ---------------------------------------------------------
##                                 log(MM)  log(Domains) log(PV)  log(Time) log(Length)
##                                    (1)       (2)       (3)      (4)       (5)
## -------------------------------------------------------------------------------------------
## I(RA == "seeking")                     0.16      0.13      0.53    1.32*     1.09
##                                 (0.35)    (0.28)    (0.41)  (0.80)    (0.78)
##
## I(UA == "seeking")                    0.72**   0.73***     0.57   1.98**    2.05***
##                                 (0.34)    (0.27)    (0.39)  (0.79)    (0.76)
##
## days_act                           0.01***   0.01***    0.004*  0.02***   0.02***
##                                (0.002)   (0.002)   (0.002)  (0.005)   (0.005)
##
## act_mm                                         1.03***
##                                                 (0.32)
##
## act_domains                                                1.77***
##                                                             (0.26)
##
## act_domains_PV                                                      2.44***
##                                                                      (0.37)
##
## act_total_time                                                      2.06***
##                                                                      (0.69)
##
## act_total_time2                                                     2.20***
##                                                                      (0.71)
##
## factor(S2)2                          -1.09***  -0.55**   -1.02*** -1.22*    -1.06
##                                 (0.29)    (0.24)    (0.36)  (0.64)    (0.66)
##
```

```
## factor(S3)2                        -0.70**    -0.40*   -1.03*** -2.19***  -2.51***
##                                 (0.29)    (0.22)    (0.32)  (0.63)    (0.65)
##
## poly(as.numeric(D1), degree = 2)1                                  -3.51**
##                                                                    (1.62)
##
## poly(as.numeric(D1), degree = 2)2                                   -0.97
##                                                                    (1.54)
##
## factor(Q1)2                                                          1.99*
##                                                                    (1.17)
##
## factor(Q1)3                                                          0.33
##                                                                    (0.62)
##
## factor(Q1)4                                                          0.40
##                                                                    (0.75)
##
## factor(Q1)5                                                          0.90
##                                                                    (1.20)
##
## factor(Q5_5)2                                                       1.37**
##                                                                    (0.67)
##
## as.numeric(Q4)                      0.04***    0.03**    0.04**   0.06*     0.05
##                                 (0.02)    (0.01)    (0.02)  (0.03)    (0.03)
##
## factor(Q5_2)2                                                       0.80**
##                                                                    (0.38)
##
## factor(Q5_3)2                                                       -2.86*
##                                                                    (1.61)
##
## factor(Q5_4)2                       -0.49     -0.39    -0.95**   -1.29    -1.60**
```

```
##                                (0.37)   (0.28)   (0.41)  (0.79)   (0.79)
##
## factor(Q6.1)true                          0.48**          2.47***  2.05***
##                                        (0.24)          (0.66)   (0.68)
##
## factor(Q11_3)2                  1.08***  0.93***          3.49***  3.11***
##                                (0.39)   (0.33)          (0.90)   (0.92)
##
## factor(Q7new)1                  -0.89***  -0.62***  -0.53  -1.87***  -1.73***
##                                (0.27)   (0.22)   (0.34)  (0.60)   (0.61)
##
## factor(Q11_2)2                                                   1.23***
##                                                                 (0.40)
##
## factor(Q6.3)true                            0.62                  0.90*
##                                           (0.38)                (0.48)
##
## factor(Q8)2                     0.59**           0.53*  1.45**   1.77***
##                                (0.27)           (0.31)  (0.60)   (0.62)
##
## factor(Q11_4)2                   0.91**   0.61*    0.88*  2.21**   2.01**
##                                (0.44)   (0.36)   (0.46)  (0.96)   (0.98)
##
## as.numeric(Q10)                                                   0.58*
##                                                                 (0.31)
##
## factor(C)low                                0.60**     0.53**
##                                            (0.28)     (0.22)
##
## factor(E)low                                                     0.88**
##                                                                 (0.39)
##
## factor(A)low                   -1.17**  -1.26***  -1.86*** -3.56***  -3.08**
##                                (0.57)   (0.46)   (0.63)  (1.26)   (1.30)
```

```
##
## factor(N)low                                 0.67     0.74   2.88**    3.27***
##                                            (0.41)    (0.58)  (1.14)    (1.18)
##
## factor(Q1)4:factor(Q5_5)2
##
##
## I(RA == "seeking")TRUE:I(UA == "seeking")  -1.76     -1.80*    -0.75  -7.89***   -7.37*
##                                  (1.37)    (1.08)    (1.59)  (2.98)    (3.08)
##
## factor(Q1)3:factor(Q5_5)2
##
##
## factor(Q1)2:factor(Q5_5)2                                              -2.20*
##                                                                        (1.29)
##
## factor(Q1)5:factor(Q5_5)2
##
##
## Constant                           -0.13     -0.98    -2.64**   -0.98    -4.42**
##                                  (0.69)    (0.68)    (1.31)  (2.16)    (2.09)
##
## ==========================================================================================
## ==========================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

4.3 Conclusion

Based on the results I conclude that risk and uncertainty attitudes have an impact of online travel related behavior. I can confirm that the hypothesised relationship exists.

Based on the results the following hypotheses are confirmed/rejected

	Hypotheses				
H1	Risk attitude	***Decreases*** the amount of	a.	Micromoments	*Rejected*
			b.	Domains	**Confirmed**
			c.	Pageviews	**Confirmed**
			d.	Time	**Confirmed**
			e.	Lenght	**Confirmed**
H2	Uncertainty attitude	***Decreases*** the amount of	a.	Micromoments	*Rejected*
			b.	Domains	**Confirmed**
			c.	Pageviews	**Confirmed**
			d.	Time	**Confirmed**
			e.	Length	**Confirmed**

5 Discussion

Behavioral data gives us an amazing capacity to understand the online behavior of the consumers as never before. Although the level of understanding of the consumers reached by the academia and practitioners is still great before the age of behavioral data, the capability of zooming into details of the real consumer behavior and the robustness of the insights are truly fascinating. Practically behavioral data empowers researchers to answer every question concerning the observable online behavior that comes to mind in a way beyond the capacity even of the consumer himself. Due to the complexity and fragmentation the consumer is unable to recall the information in such rich details as when it is recorded. Due to this fact the relationship under investigation have not been investigated before in such a detailed manner. This research confirms the finding from previous research in the area that the attitude towards uncertainty, but not risk is responsible for choosing a number of sources used for travel decision making. Our findings contribute to previous research through passive metering by investigating the time people spend on travel websites, as well as the efforts in terms of pageviews. The results from the analysis strongly confirm the hypotheses that both risk and uncertainty attitudes account for changes in time and pageviews spend desktop devices.

5.1 Implications

5.2 Contribution

This study contribute to 4 areas, it has been used as promotional material, methodologically to the area of behavioral research, to the stream of literature of decision making under risk and uncertainty, to tourism research and makes managerial implications for better online travel related marketing activities.

The findings serve as a tangible example of behavioral data employed in practice. The work been used as promotional marketing material to promote the business of Wakoopa. Insert blogs in the appendix in case blog changes address etc. The main contribution of the research is adding methodologically to the emerging stream of research in the behavioral observational research. The research explores unobservable attitudes of one's personality assessed via questionnaires in relation to a real worlds observational data. Even though it uses an established and well known analysis techniques, the novelty in data collection techniques gives significant methodological contribution for practitioners and the academia. The research contributes to the area of behavioral economics more precisely in decision under risk and uncertainty. It expands the area in relation to of online travel related behavior. It confirm the notion ones that attitude toward risk and uncertainty play a significant role in decision making concerning travel related information search and that individuals with plausible attitude toward both constructs spend less time in planning whereas those more concerned with the risk and uncertainty in general are more careful planners.

5.3 Managerial implications

5.4 Limitations & Recommendation

1. Risk and uncertainty measurement. The risk and uncertainty attitudes have been measured using the shorted possible way for measurement assessing only the general attitude of these personality traits. Arguably there could be discrepancies between the general attitude and domain specific attitude. Therefore, the experiment can be reproduced using more sophisticated measurement of risk and uncertainty attitudes such as gambles and lotteries as well as assessing perceived risks in travel related domains.

2. Classification. The classification used in the research has been used as it is provided, however it is possible to employ better classification algorithms providing better fragmentation of the travel related behavior. Therefore, for future research I suggest the use of better classification. Micro-moment, the length of a micro-moment has been set to be 5 minutes without any reference or estimation whatsoever. There is a research that suggest how a best way cut off point in order to measure user sessions can be estimated. Furthermore, we use the same cut off point of 5 minutes across desktop and mobile device. As the activity over both type of devices largely differs it could be necessary to use different cut-off points per desktop and mobile. This would also mean that the comparison between desktop and mobile is harder as it is generally different.

3. Multiple users per device. It is known that some desktop devices are shared in the households, therefore we are recording data streams from different users and assessing these data as coming from a single user. Thus, the analysis of this data can be biased as we 1) cannot distinguish the different users and 2) cannot connect properly the survey data with the behavioral data.

4. Time is inferred. The measurement cut off point is 2 seconds anything shorter than two seconds is not recorded. Crosstab browsing on desktop devices cannot be accessed. There is a measurement error on mobile devices which makes it possible to have events with time zero.

5. Selection bias. One should not overlook the possibility for selection bias as the people who agree to be tracked can be arguably different than the ones who are

more privacy conscious and would not agree to share such information even for research purposes. Selection bias states that the selected sample who agrees to participate in the research may yield different characteristics than the sample of interest drawn from the general population.

6. Programming. The program written to process and analyze the data contains several thousand lines of code. Developing such a code in an absence of a quality assurance team and/or a proper quality assurance procedures is prompt to errors.

7. Measurement. income - time -

References

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19 (6). Ieee: 716–23.

Bonn, M. a., H. L. Furr, and a. M. Susskind. 1998. "Using the Internet as a Pleasure Travel Planning Tool: an Examination of the Sociodemographic and Behavioral Characteristics Among Internet Users and Nonusers." doi:10.1177/109634809802200307.

Hastie, TJ, and D Pregibon. 1992. "Statistical Models in S, Chapter Generalized Linear Models." *Wadsworth & Brooks/Cole* 51.

Jani, Dev, Jun-Ho Jang, and Yeong-Hyeon Hwang. 2014. "Big Five Factors of Personality and Tourists' Internet Search Behavior." *Asia Pacific Journal of Tourism Research* 19 (5): 600–615. doi:10.1080/10941665.2013.773922.

Jun, S H, C A Vogt, and K J MacKay. 2007. "Relationships between Travel Information Search and Travel Product Purchase in Pretrip Contexts." *Journal of Travel Research* 45 (April): 266–74. doi:10.1177/0047287506295945.

Kim, Heejun, Zheng Xiang, and Daniel R Fesenmaier. 2015. "Use of The Internet for Trip Planning: A Generational Analysis." *Journal of Travel & Tourism Marketing* 32 (3): 276–89. doi:10.1080/10548408.2014.896765.

Mandrik, Carter A, and Yeqing Bao. 2005. "Exploring the Concept and Measurement of General Risk Aversion." *Advances in Consumer Research* 32 (32): 531–39.

Pan, Bing, and Daniel R. Fesenmaier. 2006. "Online Information Search.

Vacation Planning Process." *Annals of Tourism Research* 33 (3): 809–32. doi:10.1016/j.annals.2006.03.006.

Quintal, Vanessa Ann, Julie Anne Lee, and Geoffrey N. Soutar. 2010. "Tourists' information search: The differential impact of risk and uncertainty avoidance." *International Journal of Tourism Research* 12 (4): 321–33. doi:10.1002/jtr.753.

Ramsey, James B. 1974. "Classical Model Selection Through Specification Error Tests." *Frontiers in Econometrics*. Academic Press New York, 13–47.

Ripley, BD. 2002. "Modern Applied Sta Tistics with S." Springer-Verlag, New York.

Shapiro, Samuel Sanford, and Martin B Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52 (3/4). JSTOR: 591–611.

Stewart, Susan I, and Christine A Vogt. 1999. "A Case-Based Approach to Understanding Vacation Planning." doi:10.1080/014904099273165.

Xiang, Zheng, Vincent P. Magnini, and Daniel R. Fesenmaier. 2015. "Information technology and consumer behavior in travel and tourism: Insights from travel planning using the internet." *Journal of Retailing and Consumer Services* 22. Elsevier: 244–49. doi:10.1016/j.jretconser.2014.08.005.

Yong, An Gie, and Sean Pearce. 2013. "A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis." *Tutorials in Quantitative Methods for Psychology* 9 (2): 79–94.